Deep Learning-based Grading of Coronary Artery Disease from Coronary CT Angiography Scans

Deep-Learning basierte Klassifizierung von Koronarkrankheit in Koronar CT Angiographie Aufnahmen

Der Technischen Fakultät
der Friedrich-Alexander-Universität
Erlangen-Nürnberg

zur

Erlangung des Doktorgrades Dr.-Ing.

vorgelegt von

Felix Julius Denzinger
aus
Neustadt an der Waldnaab, Deutschland

Als Dissertation genehmigt
von der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung:	15.12.2023
Gutachter:	Prof. Dr.-Ing. habil. Andreas Maier
	Prof. Dr. Jelmer Wolterink

Abstract

One of the deadliest disease types in modern societies is coronary artery disease (CAD). It is often related to coronary artery plaques aggregating within the vessel wall obstructing the lumen causing a stenosis. To support physicians this thesis aims to answer the following research question: *Can we perform an automated CAD assessment from coronary CT angiography (CCTA) scans using machine learning (ML)?* To answer this question, we tackle three main tasks: First, we develop deep learning (DL)-based methods capable of predicting a significant stenosis degree and whether a lesion leads to a revascularization procedure. Four approaches with different characteristics are developed on this task, striving for a prior knowledge induced data representation and ML method design. Depending on whether a prior segmentation step is performed or not we reach an area under the receiver operating characteristic curve (AUC) of 0.96/0.92 for significant stenosis detection and 0.88/0.90 for the revascularization decision target with/without the segmentation. As a second task, methods to automatically determine the coronary artery disease-reporting data system (CAD-RADS) score – a patient-level CAD severity score – is developed. We leverage the best performing approach from the first task and embed it in a task-specific hierarchical architecture to aggregate single coronary subsegment features to allow a patient-level prediction. This approach is enhanced with a synergizing heuristic centerline labeling approach and auxiliary targets to reach an AUC of 0.942, 0.950 on the task of finding patients with CAD and on the task of detecting patients with an obstructive CAD respectively. With this strong performance, we tackle a third task of evaluating the clinical applicability of our CAD-RADS scoring approach. In a first step, we examine how changing some commonly altered CCTA image formation parameters influences the predictions of our approach. Here, we find that the overall performance stays on a high level, but predictions for individual patients changes. From this we conclude a need to create a more robust approach with respect to technical variation. In a second step, we develop an approach to automatically detect a norm variant of the coronaries, as out-of-domain samples may adversely impact ML-based CAD grading systems. On this task, we achieve a strong performance with an AUC of around 0.938. Additionally, we propose a quantile-based abstention approach, as an automated CAD grading system should know when a decision is better left to the human reader. Overall, this thesis concludes that – with limitations – its main research question can be answered with a "Yes". A well-performing CAD grading system was developed, but future work on robustness with respect to technical variation, the handling of anatomical outliers and explainability of the method at hand remain on the horizon.

Zusammenfassung

In der heutigen Zeit gehört die koronare Herzkrankheit (KHK) zu den tödlichsten Erkrankungen. Sie werden meist von Plaqueablagerungen verursacht, die sich in der Gefäßwand ansammeln und das Lumen verengen, wodurch eine Stenose entsteht. Um Mediziner bei der Klassifizierung von KHK zu unterstützen, soll in dieser Arbeit die folgende Forschungsfrage beantwortet werden: *Können wir KHK in Koronar CT Angiographie (KCTA) Aufnahmen mittels machine learning (ML) automatisiert klassifizieren?* Diese Fragestellung wird in drei Schritten bearbeitet: Zunächst entwickeln wir deep learning (DL)-basierte Methoden, die in der Lage sind, den Stenosegrad von Läsionen vorherzusagen und ob diese revaskularisiert wurden. Hierfür werden vier Ansätze vorgestellt, die Domänenvorwissen in die Wahl der Datenrepräsentation und das DL-Architekturdesign einbinden. Hier erreichen wir eine area under the receiver operating characteristic curve (AUC) von 0,96/0,92 für die Erkennung signifikanter Stenosen und 0,88/0,90 für die Revaskularisierungsentscheidung mit/ohne vorherigem Segmentierungsschritt. Im zweiten Schritt werden Methoden zur automatischen Bestimmung des coronary artery disease-reporting data system (CAD-RADS)-Scores entwickelt. Hierbei handelt es sich um einen KHK-Schweregrad-Score auf Patientenebene. Wir nutzen den besten Ansatz aus dem ersten Teil und betten ihn in eine aufgabenspezifische hierarchische DL-Architektur ein, um Merkmale einzelner Koronarsubsegmente zu aggregieren und eine Vorhersage auf Patientenebene zu ermöglichen. Dieser Ansatz erreicht eine AUC von 0,942 bei der Suche nach Patienten mit KHK bzw. von 0,950 bei der Erkennung von Patienten mit einer obstruktiven KHK. Mit dieser starken Performance wollen wir im dritten Schritt die klinische Anwendbarkeit unseres CAD-RADS-Scoring-Ansatzes bewerten. Zunächst untersuchen wir, wie sich die Änderung einiger häufig variierter KCTA-Rekonstruktionsparameter auf die Vorhersagen unseres Ansatzes auswirkt. Dabei stellen wir fest, dass die Gesamtperformance auf einem hohen Niveau bleibt, während sich die Vorhersagen für einzelne Patienten ändern. Daraus leiten wir die Notwendigkeit ab, einen Ansatz zu entwickeln, der robust gegenüber technischer Variation ist. Sodann entwickeln wir einen Ansatz zur automatischen Erkennung einer Normvariante der Koronarien, da sich Samples außerhalb der Trainingsdomäne negativ auf ML-basierte KHK-Bewertungssysteme auswirken können. Auch hier erreichen wir eine gute Performance mit einer AUC von 0,938. Darüber hinaus schlagen wir einen quantilbasierten Ansatz vor, der unserem Bewertungssystem die Möglichkeit einer Enthaltung einräumt, da einem automatisierten KHK-Bewertungssystem erlaubt sein sollte, die finale Befundung einem Arzt zu überlassen. Im Ergebnis ist die Antwort auf die Hauptforschungsfrage dieser Arbeit ein „Ja, aber“. Es wurde ein gut funktionierendes KHK-

Bewertungssystem entwickelt, aber es bedarf weiterer Forschung bezüglich der Robustheit des Ansatzes im Hinblick auf technische Einflüsse, des Umgangs mit Samples mit seltenen anatomischen Varianten und der Nachvollziehbarkeit der Entscheidung des KHK-Bewertungssystems.

Acknowledgement

This piece of literature is the final embodiment of a whole PhD journey, a voyage filled with moments of joy, frustration, friendship, and scientific exchange. Therefore, I would like to express my heartfelt appreciation to everyone who contributed to this in any way. First and foremost, I express my sincere gratitude to Prof. Andreas Maier, whose persistent support and guidance extended not only throughout this PhD project but also during the final stages of my bachelor's and the majority of my master's studies. Your philosophical insights and expertise are embedded within the pages of this thesis, and I am truly grateful for being able to rely on you as a steadfast pillar of support. I would also like to express my profound appreciation to Dr. Michael Sühling from Siemens Healthineers, as well as Dr. Max Schöbinger and Dr. Sebastian Faby, who enabled the realization of this thesis from the industry side. Furthermore, I am deeply indebted to my two exceptional supervisors, Dr. Michael Wels and Prof. Katharina Breininger. Dr. Wels provided continuous encouragement and uplifting feedback throughout this thesis, consistently highlighting the positive aspects of my work and offering invaluable insights. Likewise, Prof. Breininger stood beside me as a brilliant source of scientific and social feedback whenever needed. Their tireless support and guidance have been instrumental in shaping this journey. It was ultimately Prof. Breininger and Dr. Tobias Würfl who inspired me, through their enthusiasm and inexhaustible dedication in the deep learning lecture, to embark on the pursuit of a PhD. Fueled by their passion for this captivating topic, I joined the deep-learning lecture team, where I had the pleasure of meeting my first partner in crime, Florian Thamm. Together, we had an incredible time teaching numerous students and refining the deep-learning exercises. I deeply appreciate your friendship and your invaluable contributions as a scientific sparring partner throughout this journey. I also acknowledge Florian Kordon as my second partner in crime, with whom I joyfully supervised a major part of my students, forging friendship and scientific insights. Furthermore, I want to express my gratitude to Dr. Philipp Roser and Dr. Christian Marzahl. Your presence and the uplifting exchanges we shared at the lab during the challenging times of the pandemic played an integral role in preserving my sanity. I am sincerely thankful for the camaraderie we developed. When the pandemic receded, I thoroughly enjoyed the vibrant atmosphere of the lab, where the entire pattern recognition community felt like a closely-knit family. Without any

particular order, I would like to extend my gratitude to Dr. Prathmesh Mahdu, Paula Andrea Pérez-Toro, Frauke Wilm, Nora Gourmelon, Laura Pfaff, Sonja Kunzmann, Fabian Wagner, Maximilian Reymann, Maximilian Rohleder, Noah Maul, Luis Carlos Rivera Monroy, Mareike Thies, Annette Schwarz, Chang Liu, Kai Packhäuser, Fuxin Fan, and Lukas Folle. Each of you contributed to making this period of my PhD journey immensely enjoyable, with a perfect blend of social activities and captivating scientific discussions. Furthermore, I wish to thank Prof. Elmar Nöth for his role in my academic journey. Your wisdom and expertise are greatly respected and acknowledged. Moreover, I extend my heartfelt appreciation to all the students I had the privilege of supervising: Jingping Li, Mahnoor Tanveer, Jiayue Zhao, Leonhard Rist, Stephanie Mehltretter, Celia Martín Vicario, Sebastian Dörrich, Simon Langer, Klaus Fischer, Ibrahim Maniaa, Anne Edle von Querfurth, Nina Stadlbauer, and Annelie Rögele. Working with each and every one of you has not only enriched my professional growth but also allowed me to evolve as a human being. I am particularly delighted that Leonhard and Celia have joined the pattern recognition lab after completing their theses, thus becoming both friends and colleagues. Beyond the boundaries of the academic realm, I want to express my deep gratitude to my colleagues at Siemens Healthineers. First and foremost, I would like to thank Dr. Oliver Taubmann for his invaluable contributions to my research and for being a supportive friend. Additionally, I extend my thanks to Julian Anhaus, Philipp Killermann, Dr. Elisabeth Preuhs and Dr. Alexander Katzmann for their social contributions, which have greatly enriched my life as a PhD student. Outside of the academic and professional sphere, I would like to express my gratitude to my friends, who provided tenacious support throughout this journey. Although numerous individuals have influenced me, I wish to acknowledge Walter Penner for his continuous distractions, and Dr. Tim Schumacher and Christina Langer for motivating me with their own PhD stories, even though told in different research fields. Furthermore, my deepest appreciation goes to my family, which of course includes my wife's family, for their continuous support both mentally and socially throughout this endeavor. I want to convey my heartfelt gratitude to my wife, Katharina, who endured the challenges of this project alongside me, providing invaluable assistance that helped me persevere. Lastly, I express my biggest thanks to my son, Julian, whose very existence and infectious enthusiasm uplifted me through the final bits of this journey.

Felix Denzinger

Contents

PART I

Introduction and Background

CHAPTER 1

Introduction

In this introductory chapter, the importance of research in the field of computer assisted identification of coronary artery disease (CAD) is motivated, and several related terms and challenges are introduced, defined, and explained. Furthermore, contributions made to the progress of research within the scope of this thesis are listed and set into context before outlining the structure of the remaining thesis.

1.1 Motivation and Medical Background

CAD is the most frequent cause of natural death throughout modern societies [Roth 20]. Therefore, the resulting socioeconomic impact is enormous [Chau 16] and research in detecting CAD to mitigate adverse outcomes is of great importance.

1.1.1 Coronary Artery Disease

CAD usually relates to atherosclerotic plaque deposits aggregating in the vessel wall as depicted in Figure 1.1. Nowadays, it is understood as an inflammatory disease, where due to inflammation of the tissue surrounding the vasculature, immune response cells aggregate inside the vessel wall [Tion 05]. These cells then form a necrotic core consisting of mostly lipid-rich tissue. Subsequently, it undergoes fibrosis or calcification. Therefore, coronary plaques consist of either lipid-rich, fibrous or calcified tissue. The necrotic core is often bordered by a fibrous cap stabilizing the plaque. Depending on their tissue composition and structure, these plaque deposits may be prone to rupture [Maur 14]. This may lead to the plaque being released into the bloodstream, causing coagulation and thrombus formation. As these thrombi may cause myocardial infarction or strokes, this scenario is sought to be mitigated, and such plaques prone to rupture – called vulnerable plaques – should be identified as soon as possible [Maur 14]. Another aspect of plaque deposits is that these extend

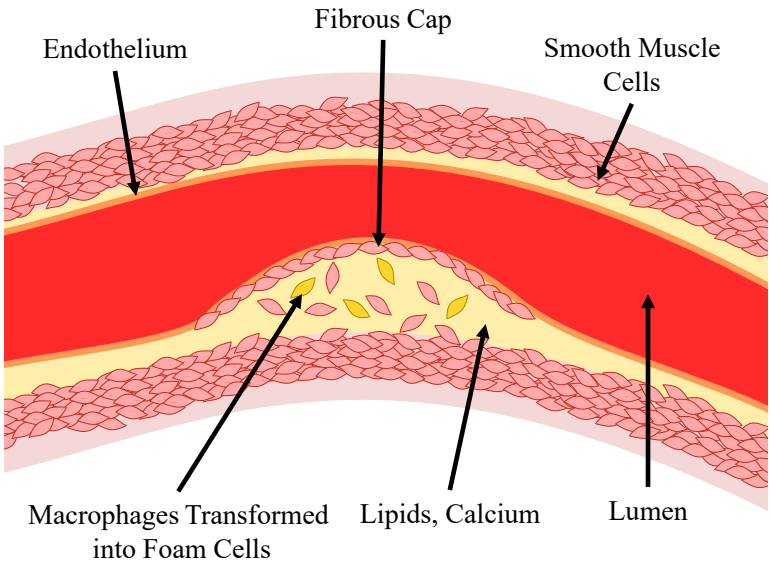

Figure 1.1: Schematic longitudinal view of an atherosclerotic plaque deposit. The lumen is narrowed due to tissue aggregation within the vessel wall. The healthy vessel wall mainly consists of smooth muscle cells. The inner wall's boundary is the endothelium. A fibrous cap stabilizes the plaque. Created using Inkscape v0.92.

into the lumen obstructing blood flow. This kind of narrowing is called stenosis, and depending on its severity, the downstream tissue, e.g., the heart muscle, may be malperfused [Cury 16]. The resulting ischemia may also cause major acute cardiac events. Consequently, stenoses severely impacting the haemodynamics, i.e. the blood flow characteristics, need immediate care. In the case of haemodynamically significant stenosis, the CAD is usually treated by minimally invasively implanting a stent at the position of the stenosis to widen the vessel. This process is called revascularization.

1.1.2 Cardiovascular Imaging and Assessment

Different modalities are considered the gold standard to identify lesions requiring action depending on whether the plaque's vulnerability or impact on the haemodynamics is of concern. For vulnerable plaques, the tissue composition and structure are most accurately assessed using intravascular ultrasound (IVUS) or optical coherence tomography angiography (OCTA) [Role 14]. These minimally invasive modalities require the insertion of a catheter usually from the femoral artery to the site of the lesion where a measuring probe either performs an ultrasound or an optical coherence tomography measurement.

Also, for assessment of potentially haemodynamically significant stenosis a catheter is used. With a pressure wire inserted through the catheter, the fractional flow re-

serve (FFR) value [Pijl 96] is measured, which is defined as the ratio of intravascular blood pressure before and after the lesion.

All modalities mentioned above require fluoroscopic guidance, i.e. continuous visualization of the catheter position using X-rays and contrast agent. Therefore, the patient is not only exposed to the risk of the invasive procedure but additionally to radiation. Furthermore, invasive assessment in the catheter laboratory is time-consuming and requires a relatively high amount of qualified personnel.

To circumvent some of these shortcomings, computed tomography (CT) as modality to assess coronaries non-invasively is becoming increasingly popular. It allows a 3D assessment of the heart and coronary vasculature. Typical scan types include the coronary calcium scoring (CCS) scan, where the heart is natively scanned, i.e. without the use of contrast agent, and reconstructed with a quantitative kernel to assess the amount of calcification present in the coronary vasculature. From this scan type, the overall calcification degree is often reported as the Agatston score (AS) [Agat 90]. It is defined as

$$\mathrm{AS} = \sum_{l \in L} s_l \tag{1.1}$$

with

$$s_l = v_l \cdot w_{\mathrm{d},l} \tag{1.2}$$

and

$$w_\mathrm{d} = \begin{cases} 0 & \max(\mathbf{C_L}) < 130\ HU \\ 1 & 130\ HU \leq \max(\mathbf{C_L}) < 200\ HU \\ 2 & 200\ HU \leq \max(\mathbf{C_L}) < 300\ HU \\ 3 & 300\ HU \leq \max(\mathbf{C_L}) < 400\ HU \\ 4 & 400\ HU \leq \max(\mathbf{C_L}) \end{cases} \tag{1.3}$$

where $l \in L$ corresponds to the singular lesion l in the set of lesions L, v_l being its volume, and w_d being a density weight factor depending on the maximum Hounsfield unit (HU) value $\max(\mathbf{C_L})$ within the segmentation of the calcification $\mathbf{C_L}$. The resulting risk groups are: no ($\mathrm{AS} = 0$), minimal ($0 < \mathrm{AS} \leq 10$), mild ($10 < \mathrm{AS} \leq 100$), moderate ($100 < \mathrm{AS} \leq 400$) and severe ($400 < \mathrm{AS}$) [Rumb 03].

A different scan type enabling a more detailed assessment of the vasculature is the coronary CT angiography (CCTA) scan, where the lumen of the vessels is enhanced by an injected contrast agent. Therefore, a CCTA scan enables lesion identification and quantification of the plaque morphology by leveraging the 3D information. This, however, comes at the cost of a lower spatial resolution compared to IVUS or OCTA. Still, several features corresponding to vulnerable plaques can be identified in CCTA scans including the so-called napkin ring sign, spotty calcifications, low attenuation

plaques, and positive remodeling [Maur 14]. Also, the degree of stenosis is an important measure, which is also highly correlated to the haemodynamic significance of the lesion. Usually, it is defined as the ratio between the measured lumen area to an estimated healthy lumen area at that vessel position. This healthy lumen area is often approximated by weighting the lumen area measurements at a healthy proximal and distal position along the centerline. However, this measurement procedure may be complicated for patients with a high calcification burden or for lesions at bifurcations where the vessel branches into two. Therefore, also eyeballing is involved when determining the stenosis degree (SD) making it a cumbersome and potentially error-prone procedure [Kiri 13]. Relevant classes of SD are no ($SD = 0\,\%$), minimal ($0\,\% < SD < 25\,\%$), mild ($25\,\% \leq SD < 50\,\%$), moderate ($50\,\% \leq SD < 70\,\%$), severe stenosis ($70\,\% \leq SD < 100\,\%$) and total occlusion ($SD = 100\,\%$) [Cury 16]. Note that stenoses with grades above and including moderate are often referred to as significant stenoses.

To have a clinical score to report the severeness of CAD on patient-level and guide through the clinical decision process, the coronary artery disease-reporting data system (CAD-RADS) score was introduced by Cury et al. [Cury 16]. It is mainly determined by propagating the most severe lesion grading to patient-level. Therefore, the same six grades as for the SD exist, with CAD-RADS 0 referring to a patient having no CAD, CAD-RADS 1 to 2 to a patient having non-obstructive CAD and 3 to 5 to a patient having obstructive CAD, which requires immediate further action. The exception of the severest stenosis degree score being propagated to patient level is the CAD-RADS grade 4. Here, CAD-RADS 4A corresponds to a severe stenosis analog to the SD, but CAD-RADS 4B is assigned if a patient has a lesion with a moderate SD in the left main segment or all three main branches have a severe stenosis. Additional modifiers for this score for non-diagnostic scans, stents, grafts, and vulnerable plaques exist but play a minor role in this thesis.

Apart from performing a measurement regarding the stenosis degree, the haemodynamic significance of coronary lesions can also be calculated by segmenting the coronary tree and computing the fluid dynamics for the entire system. This measurement is called simulated FFR [Tayl 13].

1.1.3 Coronary Artery Anatomy

To foster an understanding of the underlying anatomy, this section will briefly define the norm anatomy of the coronary arteries according to the american heart association (AHA) norm [Aust 75] and introduce some variants of this norm.

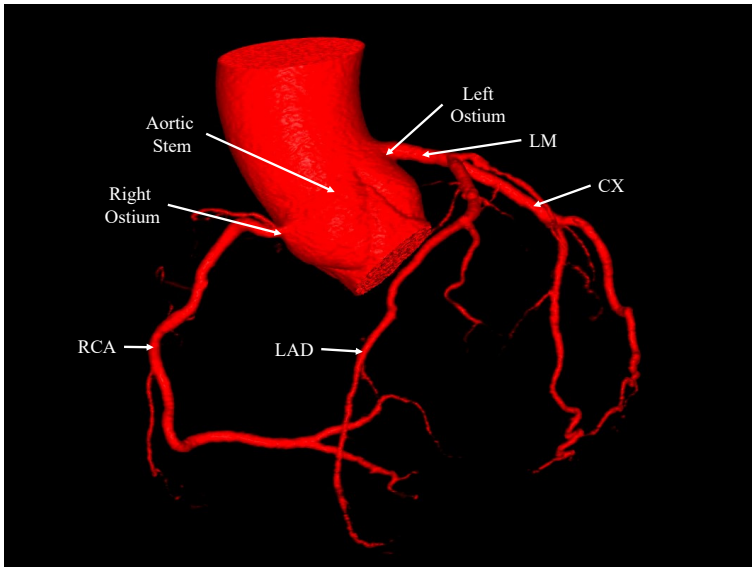

Figure 1.2: Rendering of the coronary arteries and the aortic stem. From the right ostium, the right coronary artery (RCA) arises, while from the left ostium the left main (LM) arises which subdivides into the left artery descending (LAD) and circumflex artery (CX) main branches.

A coarse overview of the coronary arteries in the form of a rendering can be seen in Figure 1.2. The aortic stem gives rise to both the RCA and LM through their respective ostium. From here, the RCA follows a path around the back of the heart and is subdivided into a proximal, middle, and distal segment with potential side branches. The distal segment may be followed by a right posterior descending artery if the vasculature is of the right or co-dominant type. On the other side of the aortic stem, the left ostium gives rise to the LM segment, which then bifurcates into the LAD and CX branch. The LAD also consists of a proximal, mid, and distal sub-segment with potential diagonal side branches and a left posterior descending artery in the case of a left or co-dominant vasculature type. Furthermore, the CX can be subdivided into sub-segments, namely a proximal, a distal, and an obtuse marginal segment.

One of the most common variants of this norm includes a trifurcation at the end of the LM segment. Hence, between the LAD and CX the ramus intermedius (RI) arises for this variant. Another norm variant of interest in this thesis is the Shepherd's crook (SC) RCA, which is defined as an RCA making a high and tortuous turn directly after the ostium. This variant is reported to complicate minimally invasive procedures in the RCA and has a prevalence of around 5 % [Shri 12]. Other norm variants and anomalies of the coronary vasculature exist but are not a main focus of this thesis.

1.1.4 Computer-Assisted Diagnosis

With the advent of increasingly potent algorithmic solutions in the field of artificial intelligence (AI) and its sub-field of deep learning (DL), these also found application in the field of medical imaging [Maie 19a]. The main question to be answered in this domain is "how can one support physicians with their increasing amount of workload using algorithms?". And the aspects in which physicians can be supported are manifold. Possible machine learning (ML) solutions in the field of CAD diagnosis include for example: tracing of the centerlines of the coronary vasculature [Zhen 13], labeling of the coronary sub-segments [Guls 14], and delineation of the inner [Luga 14] and outer vessel wall [Gros 09]. The results of these algorithms are already applied in clinical practice to facilitate the assessment from complex volumetric image data. Apart from just supporting the assessment, ML-based methods can also provide a second opinion on a patient's disease state. This kind of second reading is commonly called computer-assisted diagnosis. Note that, in theory, these kinds of algorithms could also be used as initial gatekeepers to help physicians prioritize which patients need the most care.

1.2 Contribution to the Progress of Research

Now, the general medical motivation of the research conducted in the scope of this thesis should be clear: with CCTA scans, coronary plaque deposits can be analyzed to improve risk assessment. The main research question tackled in this thesis is:

> *Can we perform an automated CAD assessment from CCTA scans using ML?*

From this main question, several subtasks emerged:

1. Finding an optimal data representation and DL architecture to predict the stenosis degree and revascularization decision for single lesions.
2. Leveraging the findings from the first step to develop task-specific DL-based approaches which directly predict the patient-wise CAD-RADS score from a hierarchical data representation.
3. Evaluation of aspects of the clinical applicability of such algorithms by examining the influence of image formation parameter variations to their predictions. Additionally, we evaluate learning with abstention, where a network does not output a class-assignment if it is uncertain of the prediction.

1.2.1 Coronary Plaque Classification

The first sub-task is tackled in two publications, which both focus on classifying a significant stenosis degree and the revascularization decision for single coronary plaque lesions. Two aspects are the focus here: optimizing the input data representation and finding a robust method embedding prior domain knowledge on the way. To enhance the input representation, it is examined whether the application of the polar transform to the cylindrical vessels yields improvements. Furthermore, several different approaches and techniques are evaluated here, including Radiomic features, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and boosting tree classifiers. In the end, taking two orthogonal longitudinal slices of the stretched lesions is found to be the data representation which yielded the best results in combination with a simple CNN. To cope with potentially sub-optimal slice selection, the concept of test time augmentation (TTA) is employed.

[Denz 19]
Section 5.2.2

F. Denzinger, M. Wels, N. Ravikumar, K. Breininger, A. Reidelshöfer, J. Eckert, M. Sühling, A. Schmermund, and A. Maier. "Coronary artery plaque characterization from CCTA scans using deep learning and radiomics". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 593–601, Springer, 2019

[Denz 20b]
Section 5.3.2

F. Denzinger, M. Wels, K. Breininger, A. Reidelshöfer, J. Eckert, M. Sühling, A. Schmermund, and A. Maier. "Deep learning algorithms for coronary artery plaque characterisation from CCTA scans". In: *Bildverarbeitung für die Medizin 2020*, pp. 193–198, Springer, 2020

1.2.2 Coronary Artery Disease Grading

After finding a suitable data representation and network architecture to classify singular plaque deposits, methods to directly predict the patient-level CAD-RADS score are sought. Here, further two publications were presented at international conferences. These also leverage domain knowledge on the anatomical definition of the coronary artery tree, which can be split in several subsegments. For each of the AHA segments, features are extracted using the representation and architecture yielding the best results on the first task. By aggregating these features on a patient-level and employing a classifier artificial neural network (ANN), we are able to accurately estimate the patient-level CAD-RADS. As a main difference between the two publications, the coronary vasculature is subdivided into segments of the same size in the

second publication. This proofs to enhance the robustness of the method and is a more straightforward labeling approach compared to the one leveraged before.

[Denz 20a] Section 6.2.2

F. Denzinger, M. Wels, K. Breininger, M. A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Sühling, and A. Maier. "Automatic CAD-RADS scoring using deep learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 45–54, Springer, 2020

[Denz 21b] Section 6.3.2

F. Denzinger, M. Wels, O. Taubmann, M. A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Suehling, and A. Maier. "CAD-RADS Scoring using Deep Learning and Task-Specific Centerline Labeling". In: *Medical Imaging with Deep Learning*, 2021

1.2.3 Clinical Application

Finally, efforts are made towards examining the clinical applicability of the CAD-RADS grading method mentioned above. Therefore, the robustness of this method regarding frequently altered reconstruction parameters is evaluated. As a summarized outcome, the overall performance of the method did not change. However, for single patients the predictions may differ for different reconstruction parameters. From this we derive the need for additional studies aiming to disentangle the biological and technical variation of the input data. Furthermore, out-of-domain samples may adversely impact the performance of the final CAD-RADS grading system. Therefore, first efforts to automatically detect infrequent norm variants are made. As a proof of concept, an approach to automatically determine the presence of SC RCA is developed. Further experiments in this work examine how to handle unsure samples, i.e. samples which human annotators cannot label confidently. Additionally, this is combined with learning with abstention, as an ML approach should be allowed to refrain from predicting if we grant the same right to human annotators. Therefore, a simple non-invasive abstention strategy is proposed.

[Denz 23a] Section 7.1.3

F. Denzinger, M. Wels, K. Breininger, O. Taubmann, A. Mühlberg, T. Allmendinger, M. A. Gülsün, M. Schöbinger, F. André, S. J. Buss, J. Görich, M. Sühling, and A. Maier. "How scan parameter choice affects deep learning-based coronary artery disease assessment from computed tomography". *Scientific Reports*, Vol. 13, No. 1, p. 2563, 2023

[Denz 23b] Section 7.2.3	F. Denzinger, M. Wels, O. Taubmann, F. Kordon, F. Wagner, S. Mehltretter, M. A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Sühling, A. Maier, and K. Breininger. "Handling Label Uncertainty on the Example of Automatic Detection of Shepherd's Crook RCA in Coronary CT Angiography". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023

1.2.4 Other Contributions

Apart from the main research area of this thesis, the author of this thesis contributed to research in other fields. These include works in the directions of standard plane regression for a set of body regions [Mart 20, Mart 22a, Mart 22b] and the spine [Doer 23], and the automatic detection of free intra-abdominal air [Taub 20]. Furthermore, clinical research in the domain of coronary plaque analysis was enabled [Denz 21a] and conducted with focus of the analysis of perivascular tissue [Mose 23]. Other contributions in the neuro CT domain include generative approaches [Tham 21] and the detection of large vessel occlusions [Tham 22]. Moreover, there were contributions to a generative approach in the histopathology domain [Kunz 22], medical image denoising leveraging known operators and being applied in both projection and image domain [Wagn 22, Wagn 23, Pfaf 23] as well as medical image segmentation [Liu 23, Quer 23], cancer survival regression [Rist 23], and strategies to migitate bias during network training [Lang 23]. Also, there were contributions to work leveraging epipolar consistency conditions for feature translation in multi-view settings [Rohl 23] and anonymizing chest radiographs [Pack 23]. These are subsequently listed in chronological order.

2020

[Taub 20]	O. Taubmann, J. Li, F. Denzinger, E. Eibenberger, F. C. Müller, M. W. Brejnebøl, and A. Maier. "Automatic detection of free intra-abdominal air in computed tomography". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 232–241, Springer, 2020

[Mart 20] C. Martín Vicario, F. Kordon, F. Denzinger, M. Weiten, S. Thomas, L. Kausch, J. Franke, H. Keil, A. Maier, and H. Kunze. "Automatic plane adjustment of orthopedic intraoperative flat panel detector CT-volumes". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 486–495, Springer, 2020

2021

[Denz 21a] F. Denzinger, M. Wels, C. Hopfgartner, J. Lu, M. Schöbinger, A. Maier, and M. Sühling. "Coronary Plaque Analysis for CT Angiography Clinical Research". In: *Bildverarbeitung für die Medizin 2021*, pp. 223–228, Springer, 2021

[Tham 21] F. Thamm, O. Taubmann, F. Denzinger, M. Jürgens, H. Ditt, and A. Maier. "SyNCCT: Synthetic Non-contrast Images of the Brain from Single-Energy Computed Tomography Angiography". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 681–690, Springer, 2021

2022

[Kunz 22] S. Kunzmann, C. Marzahl, F. Denzinger, C. Bertram, R. Klopfleisch, K. Breininger, V. Christlein, and A. Maier. "First Steps on Gamification of Lung Fluid Cells Annotations in the Flower Domain". In: *Bildverarbeitung für die Medizin 2022*, pp. 223–228, Springer, 2022

[Mart 22a] C. Martín Vicario, F. Kordon, F. Denzinger, J. S. El Barbari, M. Privalov, J. Franke, A. Maier, and H. Kunze. "Normalization techniques for CNN based analysis of surgical cone beam CT volumes". In: *Medical Imaging 2022: Image Processing*, pp. 648–652, SPIE, 2022

[Mart 22b] C. Martín Vicario, F. Kordon, F. Denzinger, J. S. El Barbari, M. Privalov, J. Franke, S. Thomas, L. Kausch, A. Maier, and H. Kunze. "Automatic plane adjustment of orthopedic intraoperative flat panel detector CT-volumes". *Journal of Medical Imaging*, Vol. 9, No. 3, pp. 034001–034001, 2022

[Tham 22] F. Thamm, O. Taubmann, M. Jürgens, A. Thamm, F. Denzinger, L. Rist, H. Ditt, and A. Maier. "Building Brains: Subvolume Recombination for Data Augmentation in Large Vessel Occlusion Detection". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 634–643, Springer, 2022

[Wagn 22] F. Wagner, M. Thies, F. Denzinger, M. Gu, M. Patwari, S. Ploner, N. Maul, L. Pfaff, Y. Huang, and A. Maier. "Trainable joint bilateral filters for enhanced prediction stability in low-dose CT". *Scientific Reports*, Vol. 12, No. 1, pp. 1–9, 2022

2023

[Quer 23] A. von Querfurth, F. Kordon, F. Denzinger, K. Breininger, and H. Kunze. "Lung, nodule and airway segmentation using partially annotated data". In: *Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling*, pp. 409–418, SPIE, 2023

[Mose 23] P. T. Moser, R. Schernthaner, C. Loewe, A. Strassl, F. Denzinger, S. Faby, M. Wels, V. Nizhnikava, K. Uyanik-Uenal, A. Zuckermann, M.-E. Stelzmüller, and D. Beitzke. "Evaluation of perivascular fat attenuation with coronary CT angiography in cardiac transplantation patients: an imaging biomarker candidate for prediction of cardiac mortality and re-transplantation". *European Radiology*, pp. 1–9, 2023

[Doer 23] S. Doerrich, F. Kordon, F. Denzinger, J. S. El Barbari, M. Privalov, S. Y. Vetter, A. Maier, and H. Kunze. "Fast 3D YOLOv3 based standard plane regression of vertebral bodies in intra-operative CBCT volumes". *Journal of Medical Imaging*, Vol. 10, No. 3, p. 034503, 2023

[Liu 23] C. Liu, L. Folle, F. Denzinger, and A. Maier. "Whole-body Multi-organ Segmentation using Distance Attention". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023

[Wagn 23] F. Wagner, M. Thies, L. Pfaff, O. Aust, S. Pechmann, D. Weidner, N. Maul, M. Rohleder, M. Gu, J. Utz, F. Denzinger, and A. Maier. "On the Benefit of Dual-domain Denoising in a Self-supervised Low-dose CT Setting". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023

[Pfaf 23] L. Pfaff, F. Wagner, J. Hossbach, L. Pfaff, E. Preuhs, N. Maul, M. Thiess, F. Denzinger, M. Nickel, T. Wuerfl, and A. Maier. "Robust Multi-Contrast MRI Denoising Using Trainable Bilateral Filters Without Noise-Free Targets". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023

[Rist 23] L. Rist, O. Taubmann, A. Mühlberg, F. Denzinger, F. Thamm, Nörenberg, J. Holch, S. Maurus, L. Gebauer, T. Huber, and A. Maier. "Spatial Lesion Graphs: Analyzing Liver Metastases with Geometric Deep Learning for Cancer Survival Regression". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023

[Lang 23] S. Langer, O. Taubmann, F. Denzinger, A. Maier, and A. Mühlberg. "DeepTechnome: Mitigating Unknown Bias in Deep Learning Based Assessment of CT Images". In: *Bildverarbeitung für die Medizin 2023: Proceedings, German Workshop on Medical Image Computing, Braunschweig*, pp. 177–182, Springer, 2023

[Rohl 23] M. Rohleder, C. Pradel, F. Wagner, M. Thies, N. Maul, F. Denzinger, A. Maier, and B. Kreher. "Enabling Geometry Aware Learning Through Differentiable Epipolar View Translation". In: *Accepted at the International Conference on Medical Image Computing and Computer-Assisted Intervention*, p. , Springer, 2023

[Pack 23] K. Packhäuser, S. Gündel, F. Thamm, F. Denzinger, and A. Maier. "Deep Learning-based Anonymization of Chest Radiographs: A Utility-preserving Measure for Patient Privacy". In: *Accepted at the International Conference on Medical Image Computing and Computer-Assisted Intervention*, p. , Springer, 2023

1.3 Thesis Structure

This thesis is structured as follows and as depicted in Figure 1.3. After introducing and motivating the underlying medical problem behind this thesis and some core terminology in this Chapter, Chapter 2 will provide a some background knowledge on the CCTA modality and CCTA in particular. Furthermore, Chapter 3 includes the theoretical foundation of the methods applied within this thesis. It starts with a general introduction of ML, continues with a limited selection of feature-based classical ML approaches, and finally elaborates on the concept behind DL and the main DL constituents. Moreover, Chapter 4 will help the reader to set the contributions of this thesis into context by providing an overview of the related work done in the field of DL-based CAD assessment. Next, as the reader should be aware of the background behind this thesis, the contributions made in the scope of this dissertation will be described regardingcoronary artery plaque classification in Chapter 5, coronary artery disease classification in Chapter 6 and clinical applicability in Chapter 7. The thesis wraps-up with an outlook in Chapter 8 and a summary in Chapter 9.

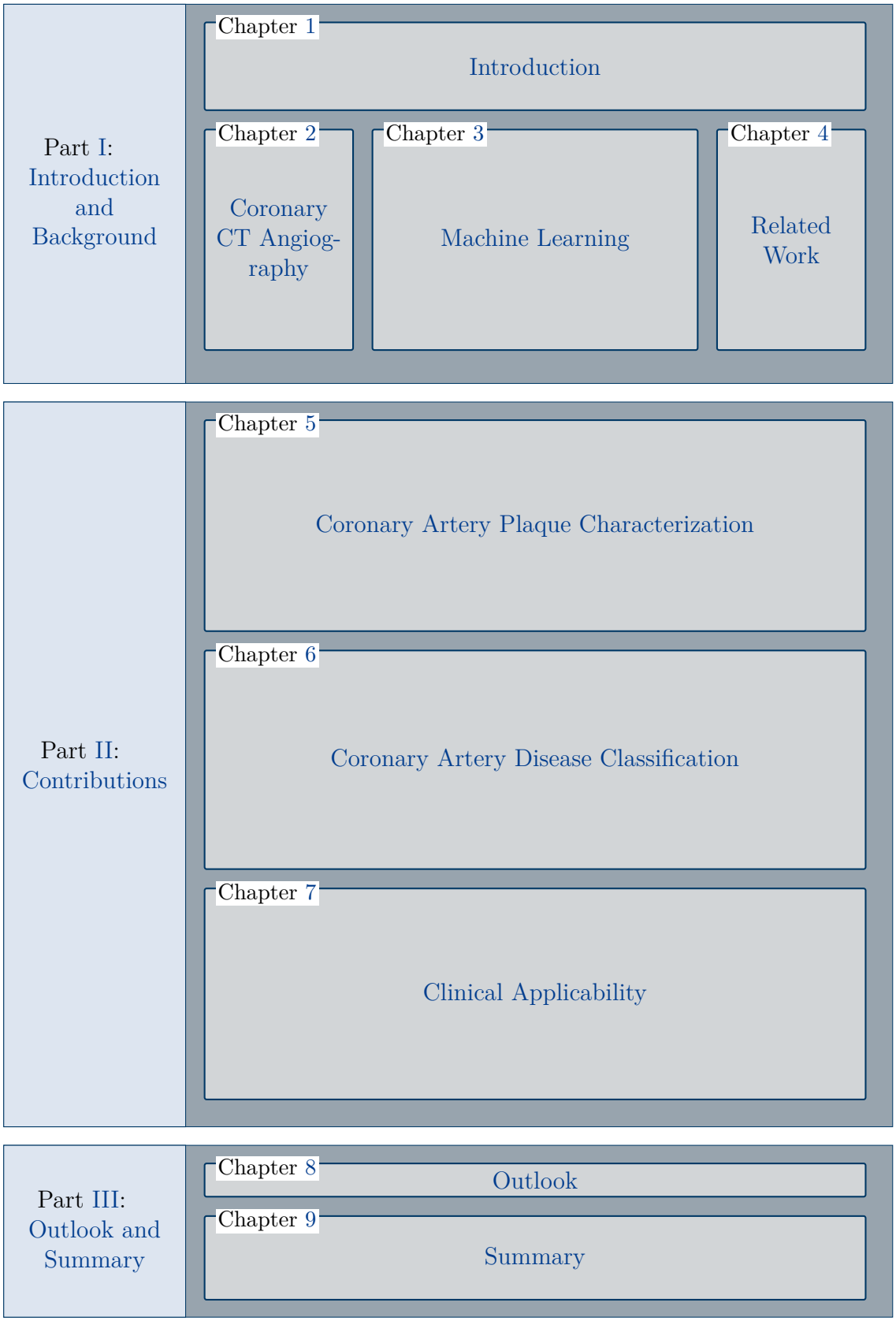

Figure 1.3: Overview over the thesis structure. The size of the individual blocks corresponds to the size of the individual chapter.

CHAPTER 2

Coronary CT Angiography

This chapter presents the computed tomography (CT) scanning type used for the research conducted in this thesis: coronary CT angiography (CCTA). First, X-rays and the CT modality are explained as these build the foundation of CCTA. Further, the CCTA acquisition type together with its unique challenges is elaborated on. Note that this chapter will only superficially cover imaging physics as these are not the focus of this thesis. Sections 2.1 and 2.2 are based on Maier et al. [Maie 18] if no other reference is given.

2.1 X-rays

An essential prerequisite before talking about CT are X-rays. This term describes ionizing electromagnetic radiation, commonly generated by an X-ray tube, where electrons are accelerated from a cathode to an anode. As the electrons hit the anode, X-rays are emitted from the anode material. Usually, the resulting X-ray cone is aimed at a patient where the X-rays may interact with the matter through absorption or scattering, which changes their direction and energy. Denser tissues and atoms with a higher atomic number lead to higher attenuation. In CT systems, the traversed X-rays hit a detector, which measures the location and number of the X-rays. For an individual X-ray beam the measured intensity I is defined by Lambert-Beer's law:

$$I = I_0 \cdot \exp\left(-\int \zeta(\mathrm{x})\mathrm{dx}\right) \tag{2.1}$$

where I_0 is the initial beam intensity and $\zeta(\mathrm{x})$ is the position-dependent attenuation coefficient. Note that this simple equation only holds for a mono-energetic X-ray as, in reality, $\zeta(\mathrm{x})$ is also dependent on the energy distribution in the used X-ray tube spectrum.

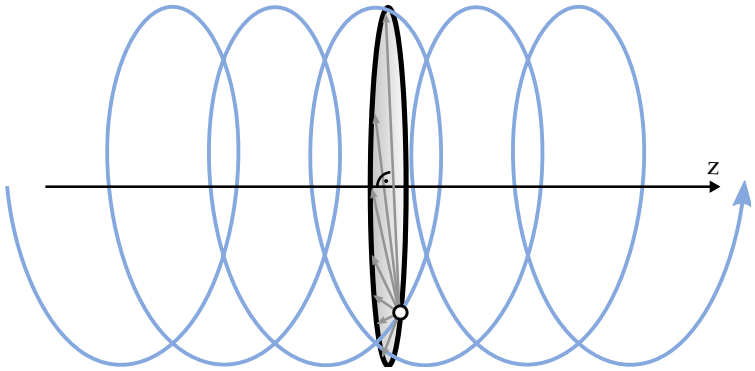

Figure 2.1: Helical trajectory of the source and detector orbiting the patient. Image taken from Maier et al. [Maie 18] under CC BY 4.0 license.

2.2 Computed Tomography

Using an X-ray tube and detector pair rotating around a patient, projection images can be measured. Usually, for CT imaging outside of the intra-operative suite, the trajectory of the source and detector pair is a helix created by altering the angle θ and moving the patient on a table with a pitch in z-direction as visualized in Figure 2.1. Pitch is defined as the distance traveled during one 360 ° rotation divided by the utilized detector width in z-direction. For each position on this trajectory, a projection image is acquired. This step is also called forward projection. Usually, the filtered back projection (FBP) algorithm is applied to reconstruct a CT image from these measurements. One step of it is a back-projection of the data, i.e. the data is smeared back into the volume to be reconstructed. However, the back projection is not the exact inverse of the forward projection. The reasoning for this can be deducted from the Fourier slice theorem as depicted in Figure 2.2. For each angle θ a projection is obtained. If only one slice is considered, this projected line $p_\theta(\mathrm{s})$ corresponds to one line in the Fourier space $F(\mathrm{u}, \mathrm{v})$ (u and v describe the coordinates in Fourier space) at angle θ. If a sufficient number of viewing angles is sampled, a representation of the object in Fourier space can be acquired. With this, a spatial slice can be reconstructed using the inverse Fourier transform to obtain the object to be reconstructed, which is described by the function $f(\mathrm{x}, \mathrm{y})$.

However, this sort of sampling neglects that the 1D Fourier transform leveraged to bridge the gap between $p_\theta(\mathrm{s})$ and the corresponding line in $F(\mathrm{u}, \mathrm{v})$ is defined as:

$$P(\xi, \theta) = \int_{-\infty}^{\infty} p_\theta(s) e^{-2\pi i \xi \mathrm{s}} \mathrm{ds} \tag{2.2}$$

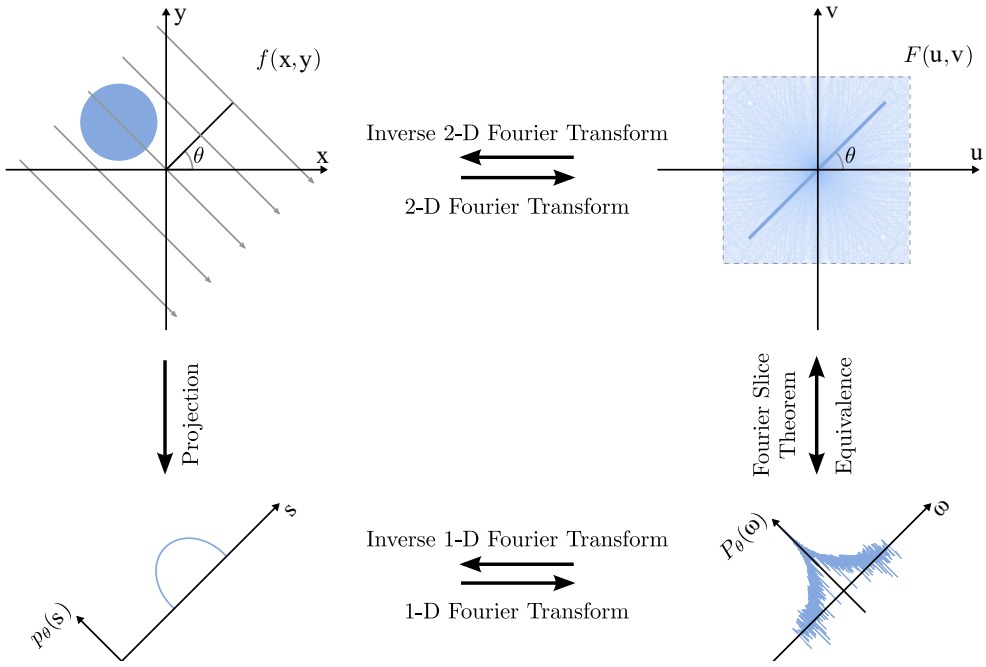

Figure 2.2: Fourier slice theorem. There is an equivalence between a projected line $p_\theta(\mathrm{s})$ for angle θ, it's 1D Fourier transform, and a line in Fourier space at angle θ. Image taken from Maier et al. [Maie 18] under CC BY 4.0 license.

and the additional parameter ξ needs to be considered as it leads to an oversampling in the center of Fourier space. Therefore, the formula to compute the image function $f(\mathrm{x},\mathrm{y})$ using FBP is given as:

$$f(\mathrm{x},\mathrm{y}) = \int_0^\pi p_\theta(\mathrm{s})h(\mathrm{s})|_{s=\mathrm{x}\cos\theta+\mathrm{y}\cos\theta}\mathrm{d}\theta. \tag{2.3}$$

Here, $h(\mathrm{s})$ refers to the inverse Fourier transform of $|\xi|$. It is also called ramp-filter and suppresses the over-sampled low frequencies while enhancing the high frequencies as depicted in Figure 2.3. Popular filter embodiments include the Ram-Lak and Shepp-Logan filters. The Ram-Lak filter is defined as

$$h(s) = \frac{\mathrm{sinc}(\frac{s}{\Delta s})}{2(\Delta s)^2} - \frac{\mathrm{sinc}^2(\frac{s}{2\Delta s})}{4(\Delta s)^2} \tag{2.4}$$

with Δ denoting the laplace-operator. By altering the width of the kernel, the amount of noise and sharpness can be adapted. The resulting volume consists of attenuation coefficients which are commonly normalized to Hounsfield unit (HU) values where water has a HU value of zero and air a value of -1000.

An extension of the original FBP algorithm, which is of interest in this thesis, is the advanced modeled iterative reconstruction (ADMIRE) method [Rami 18]. As the exact implementation is proprietary to Siemens Healthcare GmbH, only a high-level overview will be given here. During a single FBP step, individual projections are

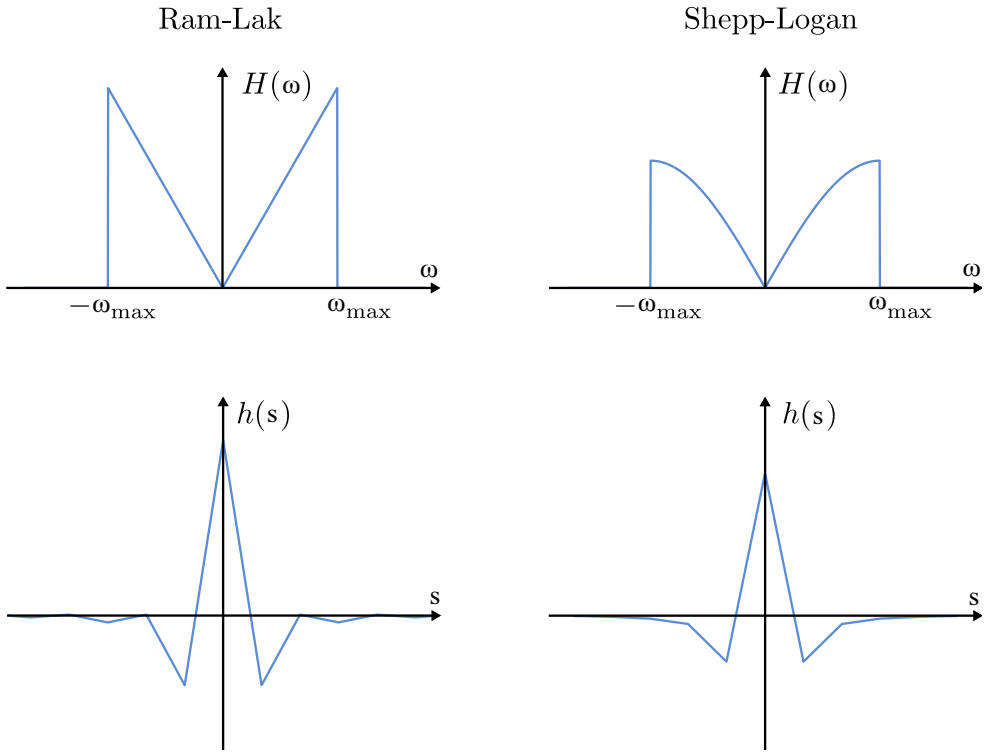

Figure 2.3: Two commonly used filter kernels for FBP – the Ram-Lak and Shepp-Logan filter – in frequency (top) and spatial (bottom) domain. Note that the width of the filter can be altered by adapting ω_{min} and ω_{max} to obtain different image characteristics. Image taken from Maier et al. [Maie 18] under CC BY 4.0 license.

statistically weighted using a model-based approach and a region around each voxel is considered to distinguish anatomical information from image noise with the latter being suppressed. Furthermore, the scanner geometry and so-called flying focal spot are leveraged to reduce the amount of streaking and windmill artifacts. The result of a single FBP iteration with said enhancements is then validated against the original projection data. The statistical weighting is then updated and used for another FBP iteration. This procedure is then repeated up to 5 times in practice, and the number of iterations is branded as "ADMIRE strength". With each iteration the amount of noise in the reconstructed volume is reduced.

2.3 Coronary Computed Tomography Angiography

As the objective of CCTA scans is to enhance the coronary vasculature, these are scanned while administrating contrast agent. This contrast agent usually is iodine-based as iodine is a bio-compatible substance with a relatively high atomic number. This leads to a higher attenuation of X-rays hitting blood mixed with the contrast

agent and vessels appearing brighter than for native scans without contrast agent. As a result, the vessels can be easier distinguished from their surrounding tissue.

However, one core challenge when acquiring CCTA scans is the constant motion of the heart. Usually, it is solved by simultaneously measuring an electrocardiogram (ECG) while scanning and reconstructing only with retrospectively selected projection images from the same heart phase. However, this forces the acquisition to last for multiple heart cycles, which hence increases the influence of breathing motion on the reconstruction. As a consequence and as the heart usually cannot be scanned with a single gantry rotation so-called stacking artifacts may occur. As the volume consists of a set of overlapping sub-volumes from different heart cycles the breathing motion between the sub-volumes may introduce an offset in the xy-plane. Note that also motion in the z-direction occurs, which can only partly be suppressed when mixing overlapping sub-volumes.

CHAPTER 3

Machine Learning

This chapter gives an overview of machine learning (ML) in general and explains the most relevant sub-areas for this thesis in more detail. The first part features a general introduction to the high-level concept of ML and defines several key terms and challenges. Next, some constituents of classical ML relevant to this thesis are explained. Finally, the concept behind deep learning (DL) will be elaborated on together with the most crucial building blocks used in this thesis.

3.1 Introduction

In a nutshell, ML algorithms leverage features $\mathbf{x}$ extracted from measurements $\mathbf{m}$ of real-world information $\boldsymbol{\omega}$ to assign a sample to a class in the form of a prediction $\hat{y}$. This general procedure is depicted as the pattern recognition pipeline in Figure 3.1 [Niem 13]. Measurements $\mathbf{m}$ can be obtained from various different acquisition devices, with the aim of digitizing information $\boldsymbol{\omega}$ from the real world $\boldsymbol{\Omega}$. These measurements are usually not directly processed by an ML algorithm. Hence, they are further pre-processed and a set of features $\mathbf{x}$ is extracted from each sample. Commonly used pre-processing includes denoising, normalization, or dimensionality reduction. The resulting feature vector $\mathbf{x}$ is then processed by an ML algorithm, which in the most simple case may just be linear thresholding of a single feature. Usually, the parameters $\boldsymbol{\alpha}$ of this so-called classifier are determined during a training phase and then applied in a test phase. The output of the classifier is a prediction $\hat{y}$, which assigns the sample to a class. For measuring the performance and for optimization of the classifier, the predictions are usually compared to a ground truth y. This ground truth is commonly created by human annotators.

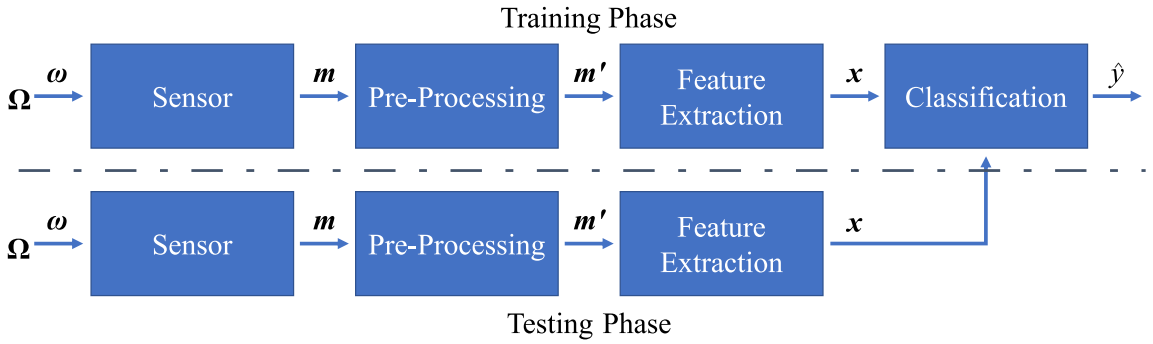

Figure 3.1: Flow of the pattern recognition pipeline. Information $\boldsymbol{\omega}$ from the real world $\boldsymbol{\Omega}$ are measured using a sensor to obtain measurements $\mathbf{m}$. Next, measurements are pre-processed to obtain $\mathbf{m'}$ from which features $\mathbf{x}$ are extracted. These are then processed by a classifier to obtain a prediction $\hat{y}$. The parameters of the classifier are set during the training phase and applied during the test phase.

Instead of predicting a sample's membership to one or multiple classes, a continuous output can be obtained as well by predicting a sample's membership to one of infinitely many ordered classes. This is called regression. The term ordinal classification describes a mix of classification and regression, where there is a finite set of ordered classes to which a sample can be assigned. An example of this are medical severeness grades. Furthermore, for a single sample, more than a single target can exist. This is the case for the task of segmentation, where for every single pixel or voxel in an image its respective class membership is predicted.

As a practical example, the biological information $\boldsymbol{\omega}_{\text{bio}}$ of a patient is measured with a computed tomography (CT) scanner under the use of contrast agent to obtain projection images of the coronaries. These then get pre-processed by a filtered back projection (FBP)-based algorithm [Maie 18] to obtain a coronary CT angiography (CCTA) volume (c.f. Chapter 2). Note that each of these initial steps may already introduce some inaccuracies into the system. For once, the projection images exhibit noise, and the spatial size of individual voxels impacts how accurately $\boldsymbol{\omega}_{\text{bio}}$ can be measured. As an exemplary task to solve one might seek to design an ML algorithm to detect dextrocardia, which is defined as the heart being located on the right side of the thorax instead of the usual left side [Van 64]. With an additional pre-processing step, the voxels of the vasculature can be separated from the rest by using the prior knowledge that iodine-enriched blood has a specific range of Hounsfield unit (HU) values. Since the coronary vasculature wraps around the heart, the orientation and position of the vasculature, which can be computed by first- and second-order moments in the x-direction, can be used as a feature. This feature can then be used as an input to a classifier that decides whether a patient has dextrocardia or not by simply looking at whether this direction vector points to the left or right of the

patient. This process of designing an algorithm can be described as model-driven. One knows that in the case of dextrocardia the heart points to the right side, which is exploited by a hand-crafted feature.

A contrary philosophy to approach the pattern recognition pipeline is purely data-driven, combining the feature extraction and classification step. The increasingly popular field of DL leverages exactly such a data-driven mechanism. Here, a so-called artificial neural network (ANN) – composed of a sequence of functions with optimizable parameters – learns to extract features from the pre-processed measurements $\mathbf{m}'$ and also acts as a classifier. In this case, a multitude of data is shown to an algorithm to optimize its parameters, e.g., CCTA scans of patients with or without dextrocardia. But for dextrocardia – an infrequent condition, which only around 1 out of 12.000 persons are born with [Bohu 07] – an algorithm has to be presented a representative set of both healthy and diseased patients to be able to solve this task. In the field of computer vision (CV) where data is usually available in far larger quantities – e.g., the popular ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset contains millions of images with thousands of samples [Russ 15] – this sort of procedure leads to impressive performance improvements over traditional ML. However, in the medical domain, the number of available samples for individual tasks is usually sparse and pathological cases are often in the minority. Additional problems inherent in the medical domain are that some pathologies are hard to detect – even for trained physicians – leading to a relatively high amount of inter- and intra-observer variability and that each clinical site has its own "best practices" regarding how data is acquired and interpreted.

In general, DL-based approaches often outperform classical approaches if a sufficient quantity of data of all classes to be detected is given. If this condition is not fulfilled, combining the best of both worlds often yields good results. For example, Maier et al. showed that known operators – prior knowledge enriched components embedded into an ANN – reduce the error bound [Maie 19b]. Other components to optimize in a model-driven fashion are, for example, the data representation used as an input for the classifier. As a rule of thumb, one should try to have a representation that only differs with regard to the information about the healthy and pathological aspects of the data and have a data collection that includes all possible variations of these aspects. Redundant or unimportant information in the data representation may threaten the algorithms' generalizability, as data-driven approaches tend to look for shortcuts. E.g., in a case where ill patients have wires attached to them visible in a CT scan, an ML approach might learn the probability of the patient exhibiting a pathology is increased based on the presence of wires. However, this so-called con-

founder might not be given at a different hospital with a different scan or workflow procedure. In summary, some key challenges in the domain of DL for medical image analysis are the sparsity of data, class imbalances, generalizability across sites, finding an optimal data representation, and detecting and preventing confounders from leading to shortcuts.

3.2 Classical Approaches

Now that ML in general and some inherent key challenges are defined some classical concepts used in this thesis are explained in this section. First, a set of features relevant for this thesis is defined, which are shape-based features. Furthermore, the eXtreme Gradient Boosting (XGBoost) algorithm – a classifier characterized by a usually good performance and fast optimization – is eleborated on.

3.2.1 Shape-Based Features

As described in Section 3.1, a key part of designing a classical ML approach is the selection of an appropriate feature representation. In case of the coronary vasculature a set of hand-crafted features are an obvious choice when trying to assess the shape of a stenotic vessel: shape-based features. These form a sub-group in the field of Radiomics [Lamb 12]. To be able to calculate them a prior segmentation of the lesion or sample to be analyzed is required, which on the one hand might be an additional error inducing step. However, the resulting features themselves are invariant regarding intensity shifts. Commonly used shape-based features include: the volume, surface area, sphericity, spherical disproportion, eigenvectors and eigenvalues. The reader is pointed to Van Griethuysen et al. [Van 17] for exact definition of these features.

3.2.2 Decision Trees

Decision trees form a quite simple sort of classifiers. An example is given in Figure 3.2. The nodes of the tree correspond to single features. For each single feature at each node a linear decision boundary is applied. Finally, the deepest nodes correspond to the predictions of the classifier. When designing a decision tree classifier, the depth of the tree is the main parameter to set. Advantages include that they are relatively easy to interpret, can handle continuous and categorical data, and are invariant to scale and transformations. However, they may include irrelevant features, which hinder interpretability, are not applicable for more complex problems. For example it struggles to model the XOR problem [Hast 09].

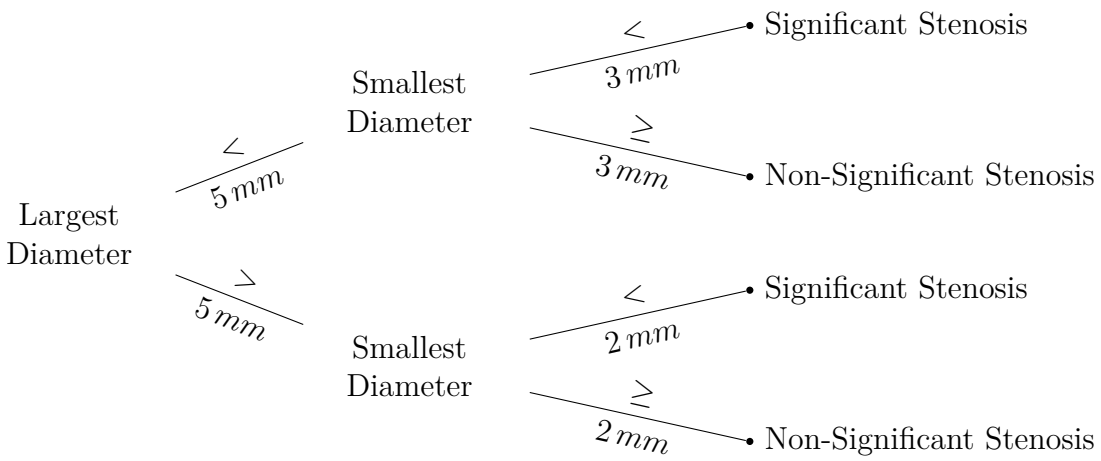

Figure 3.2: Example of a decision tree. Each feature is a node of the tree. By thresholding a single sample is assigned to a subbranch of the tree. The final nodes of the tree are the predictions.

3.2.3 XGBoost

An extension of basic decision trees was proposed by Chen et al. [Chen 16] and branded XGBoost. The final prediction of the XGBoost algorithm is made by having a majority vote of multiple singular decision trees. In order to create this set of decision trees the following loss function is minimized:

$$\mathcal{L}_{\text{XGB}} = \sum_{i=1}^{n} \mathcal{L}_{\text{diff}}(y_i, \hat{y}^{(t-1)} + g_t(\mathbf{x}_i)) + \gamma(g_t) \tag{3.1a}$$

$$\text{where } \gamma(g) = w_\tau \tau + \frac{1}{2} w_{\text{reg}} ||\boldsymbol{\alpha}||^2 \tag{3.1b}$$

Here, a new tree $g_t(\mathbf{x}_i)$ is searched in a greedy manner and added to the set of decision trees of such that the set minimizes $\mathcal{L}_{\text{XGB}}$. This step is performed n times. A main constituent is a convex differentiable loss function $\mathcal{L}_{\text{diff}}$ computed using label y_i and prediction $\hat{y}^t$ with t denoting the training step. Every single tree consists of τ leafs with corresponding leaf weights w. These parameters are considered in a regularization term $\gamma(f)$, which seeks to prevent overfitting by penalizing large values of τ and $\boldsymbol{\alpha}$ with corresponding weighting factors w_τ and w_{reg}. To obtain optimal weights $\boldsymbol{\alpha}^*$, $\mathcal{L}_{\text{XGB}}$ can be optimized as it is convex and differentiable. XGBoost tackles some problems inherent in decision trees (e.g., the usually low model capacity) and has inherent mechanisms to prevent overfitting. Hence it is a common first choice when dealing with data consisting of handcrafted feature vectors.

3.3 Deep Learning

Extending on the initial introduction of DL and ANNs in Section 3.1, this section aims to lay the knowledge foundation regarding the inner workings of ANNs. First, the general principles of DL, including the composition of ANNs, some of their common characteristics, and how they are optimized, are explained. Next, the building blocks leveraged for the research in this thesis are explained, which includes the fully connected layer, common activation functions, batch normalization as well as convolutional, pooling, and recurrent layers.

3.3.1 Principles

The human brain's neural network is made of complex structures of interconnected neurons. These neurons communicate with each other through electrical and chemical signals. For an individual neuron, signals arriving from multiple other neurons aggregate until a threshold is passed, leading to the neuron firing and transmitting a signal to all subsequently connected neurons. Connections between the individual neurons, called synapses, are adapting their strength based on experience. This process, known as synaptic plasticity, enables the core brain functionalities of memory and learning [Llin 88].
Inspired by neurons, the Rosenblatt perceptron was proposed in 1958 [Rose 58]. It models a single neuron, where multiple input signals are weighted and summed up. Its decision rule is described as:

$$\hat{y} = (1 + \text{sign}(\mathbf{w}^T \mathbf{x}))/2. \tag{3.2}$$

Here, $\mathbf{w}$ corresponds to weights for each input feature x and this function assigns samples to a positive or negative class depending on the sign of the weighted sum. As this approach leads to a linear decision boundary, it cannot solve more complex tasks like modeling the XOR function. Therefore, the concept is extended with the introduction of so-called hidden layers between the input and output layer leading to a multi-layer perceptron [Rume 86]. Usually fully connected layers, also known as dense layers, are used as hidden layer. They apply an affine transformation defined as:

$$f_{\text{FC}}(\mathbf{x}) = \mathbf{W}^\top \mathbf{x} + b \tag{3.3}$$

with $\mathbf{x}$ denoting either the initial feature input or a subsequent feature representation, $\mathbf{W}$ being a matrix of size $m \times n$ with m being the number of input and n the amount of output nodes and b being an additive bias.

From a mathematical and more general perspective, a ANN can be described as a chain of functions. As an example $f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$ would describe a three layer ANN consisting of single functions f [Good 16]. In practice, scaling operators, e.g., a matrix multiplication with weights, are alternated with non-linear activation functions. This enables the modeling of any arbitrary function given a large enough single hidden layer and a locally bounded piece-wise activation function [Cybe 92]. However, multiple hidden layers are used in practice to model the classifier function more efficiently [Good 16].

For an ANN to be a potent classifier, its parameters need to be optimal. Typically, these are initialized randomly. As all layers usually describe derivable functions, the ANN's parameters can be optimized by calculating the gradient for each individual weight with respect to their influence on the final prediction. Four steps are commonly required to optimize a neural network: First, an input $\mathbf{x}$ is fed into the ANN, resulting in a prediction after the chain of functions is calculated. This step is called forward pass. Then a loss regarding some optimization criterion, e.g., based on the distance to a ground-truth value, is computed. Common loss functions include the cross entry loss for classification defined as

$$\mathcal{L}_{\text{BCE}} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{3.4}$$

for the binary case and the mean squared error (MSE) loss for regression tasks, defined as

$$\mathcal{L}_{\text{MSE}} = ||y - \hat{y}||_2^2. \tag{3.5}$$

Thirdly, the error is propagated backwards through the network by calculating the gradient for each respective layer input and all the weights. An example for this is depicted in Figure 3.3. As the forward pass is a chain of functions, this backward pass boils down to the chain rule of differentiation when it comes to the gradient with respect to the inputs. In order to calculate the gradient for each individual layer, the gradient of all latter layers can be reused. Additionally, the gradients with respect to the weights of each layer is calculated, which also requires the inputs of the forward pass. Finally, the weights of the ANN are updated based on their gradient and a gradient descent algorithm. In the simplest case stochastic gradient descent is used with

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \nu \nabla \mathcal{L}(\mathbf{w}^{(t)}, \mathbf{x}, \mathbf{y}) \tag{3.6}$$

as an update rule, with $\nabla \mathcal{L}(\mathbf{w}^{(t)}, \mathbf{x}, \mathbf{y})$ being the gradient with respect to the loss function and ν the learning rate. The most common optimizer in nowadays practice is the Adam optimizer, which also incorporates "individual learning rates" for single

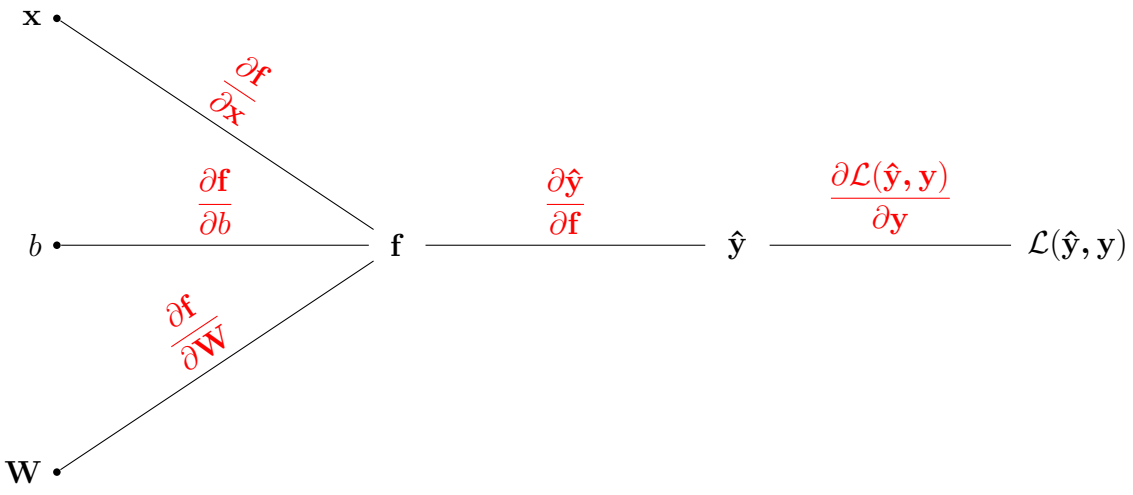

Figure 3.3: Backpropagation for a fully connected layer. In the forward pass $\mathbf{x}$, b and $\mathbf{W}$ get combined to form $\mathbf{f}$, which is subsequently processed by an activation function to receive a prediction $\hat{\mathbf{y}}$. Meanwhile in the backward pass, the partial derivatives regarding each individual contributing factor is calculated (depicted in red here). Note that the derivative with respect to the input $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$ gets passed to the previous layer, where the same calculations are performed.

weights by scaling the weight update by how much a weight was already updated in previous steps.

It features a momentum term

$$\mathbf{p^{(t)}} = \beta_p \mathbf{p^{(t-1)}} + (1 - \beta_p) \nabla \mathcal{L}(\mathbf{w}^{(t)}, \mathbf{x}, \mathbf{y}) \tag{3.7}$$

and a velocity term

$$\mathbf{v^{(t)}} = \beta_v * \mathbf{v^{(t-1)}} + (1 - \beta_v)(\nabla \mathcal{L}(\mathbf{w}^{(t)}, \mathbf{x}, \mathbf{y}) \odot \nabla L(\mathbf{w}^{(t)}, \mathbf{x}, \mathbf{y})) \tag{3.8}$$

with β_m and β_v being weighting factors and $\odot$ denoting an element-wise multiplication. These terms are then used to form the update

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \nu \frac{\mathbf{p^{(t)}}}{\sqrt{\mathbf{v^{(t)}}}}. \tag{3.9}$$

Usually, multiple samples are combined in a batch fed to the ANN at once to stabilize training.

3.3.2 Building Blocks

In the following, common building blocks of ANNs are defined. If not stated otherwise, this section is based on Goodfellow et al. [Good 16].

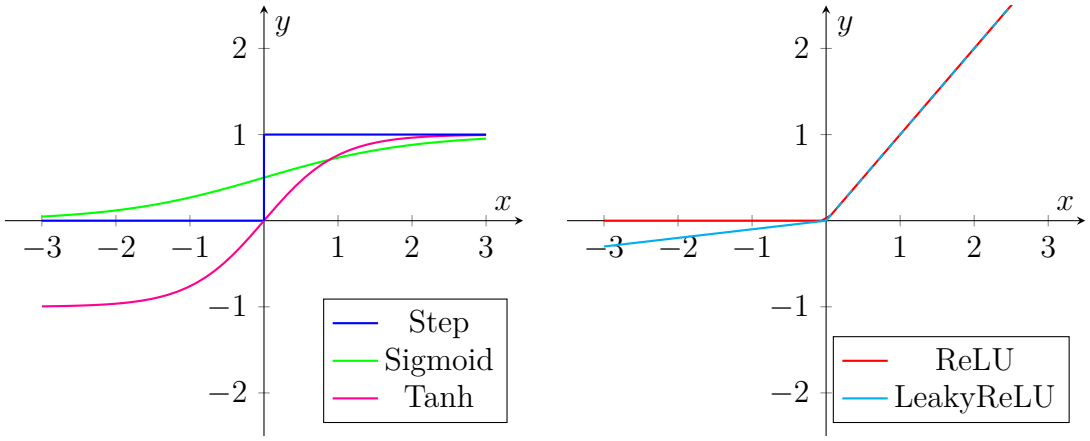

Figure 3.4: Exemplary activation functions: sigmoid and the tanh are smooth and differentiable approximations of the step function within the range of [0,1] and [-1,1] respectively. The rectified linear unit (ReLU) function is defined as $\max(0, \mathbf{x})$, is linear for positive values while setting all negative values to zero. The Leaky ReLU only scales down negative values.

Activation Functions

Usually, any sort of affine transformation layer, like the fully connected layer which was introduced before, is followed by an activation function. These functions need to be non-linear in order to model non-linearities between the input and output. Examples are given in Figure 3.4.

In the original Rosenblatt perceptron, the step function was used as an activation function. However, the step function cannot be differentiated in a useful manner for backpropagation. Therefore, differentiable approximations of the step functions – the sigmoid or tanh function – are often applied especially in the final layer of a classifier. Even though they are non-linear, they lead to the so-called vanishing gradient problem, as small ranges of the output $\mathbf{y}$ get mapped to a large range of $\mathbf{x}$ values. An activation function not negatively impacted by this effect is the ReLU function $\max(0, \mathbf{x})$, being linear for all positive and zero for all negative $\mathbf{x}$ values. However, this strict behaviour of mapping all negative to one value does lead to the fact that no gradient flows into the direction of a negative activation. This leads to ReLUs potentially "dying" and therefore not contributing to network training any longer. To overcome this initially, the bias b of fully connected layers is usually set to a small positive value. Moreover, an extension of the ReLU – the so-called Leaky ReLU – was introduced [Maas 13], which is defined as:

$$\mathcal{A}(\mathbf{x}) = \begin{cases} x \text{ for } x \geq 0 \\ 0.01x \text{ for } x < 0 \end{cases} \tag{3.10}$$

scaling down all negative values of $\mathbf{x}$ with a small factor.

Batch Normalization

A theoretical drawback of the use of ReLU non-linearities is the so-called covariate shift. As the ReLU only emphasizes positive values, the value distribution of features increases throughout the network. Also, if the value range of one layer changes, all subsequent layers need to adapt accordingly. To prevent such behaviour the concept of batch normalization layers was introduced [Ioff 15]. It enforces a zero mean and a standard deviation of one for each batch of data fed to the network. This is achieved by calculating

$$f_{\mathrm{BN}}(x) = \frac{x - \mu_x}{\sigma_x} \tag{3.11}$$

for each individual feature/activation x. The mean μ_x and standard deviation σ_x is calculated for each individual batch and a weighted moving average is collected over the whole training data collection. Additionally, a scale and shift step is introduced which enforces the batch normalization to create an identity transform. For this, the output of the normalization step $f_{\mathrm{BN}}(x)$ for each individual activation is scaled and shifted with respective trainable parameters. So far, batch normalization for the case of fully connected layers was described. For convolutional layers, which will be described next, the concept can be applied anologeously by replacing single features with feature maps.

Convolutional Layers

In the domain of image analysis, the use of fully connected layers to directly process images has several serious drawbacks. On the one hand, the information of spatial neighborhood of individual pixels is lost, which leads to a non shift-invariant classifier. On the other hand, the number of weights needed to map an image to a function rises exponentially with the image size. To overcome these problems, convolutional layers were introduced [LeCu 89b, LeCu 89a]. For an input image $\mathbf{I}$ a 2D convolutional layer is defined as

$$f_{\mathrm{Conv}}(\mathrm{x}, \mathrm{y}) = (\mathbf{I} * \mathbf{K})(\mathrm{x}, \mathrm{y}) = b + \sum_m \sum_n \mathbf{I}(\mathrm{x}, \mathrm{y})\mathbf{K}(\mathrm{x} - m, \mathrm{y} - n) \tag{3.12}$$

at position (x, y), with $\mathbf{K}$ being a kernel of size $M \times N$ and a bias b. Again, it is usually followed by a non-linear activation function. This mathematical operation takes an input image – or feature map – and slides a filter kernel over it. At each filter position the dot product between the input values and the filter kernel is calculated. With this a new feature map is created. In practice usually multiple filters are applied at the same stage to be able to model a variety of features. Also, for efficiency, several feature maps or input channels are often processed by the same kernel across the channel

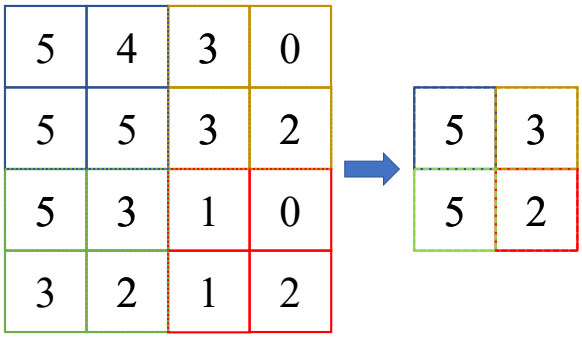

Figure 3.5: Visualization of a max pooling operation with a kernel size of 2×2 with a stride of 2.

dimension to also combine individual feature maps. In classical image processing, this operation is used for edge detection, smoothing or feature extraction by initializing the parameters of the kernel in a respective structured manner. For ANNs on the other hand, weights are initialized randomly and such sort of feature extraction capabilities are learned in a data driven fashion. The extracted features within one individual layer only combine a local part of the image. Therefore, multiple convolutional layers are often applied consecutively to also be able to model long distance dependencies.

An alternative to model long distance dependencies are dilated convolutions. For these the convolution kernel size gets increased by a fixed dilation factor $d \in \mathbb{N}$ while keeping the same amount of parameters by adding columns and rows filled with zeros between the existing parameters. With these features corresponding to lower frequencies can be extracted. Another special kind of convolutional layers, are depthwise separable convolutions, which have a filter size of 1×1 [Chol 17]. These are commonly used to combine feature maps across the channel dimension.

Pooling Layers

To reduce the spatial dimensions of feature maps so-called pooling layers are often employed. These leverage the fact that neighboring features tend to contain similar information, and allow subsequent convolutional layers to model more complex hierarchical structures. Furthermore, small shifts of the image data are compensated. Pooling layers are subdivided into local and global pooling or maximum and average pooling. The most frequently used kind is a local maximum pooling operation. It is defined by a kernel size and stride used to slide over the image (cf. Figure 3.5). Within the kernel at each position the maximum value is propagated to the next layer. Usually, the stride and kernel size have the same value to allow for a dimensionality reduction and full sampling of the image. An alternative to this is fractional max pooling [Grah 14], which reduces the spatial dimension by a factor of e.g., $\sqrt{2}$

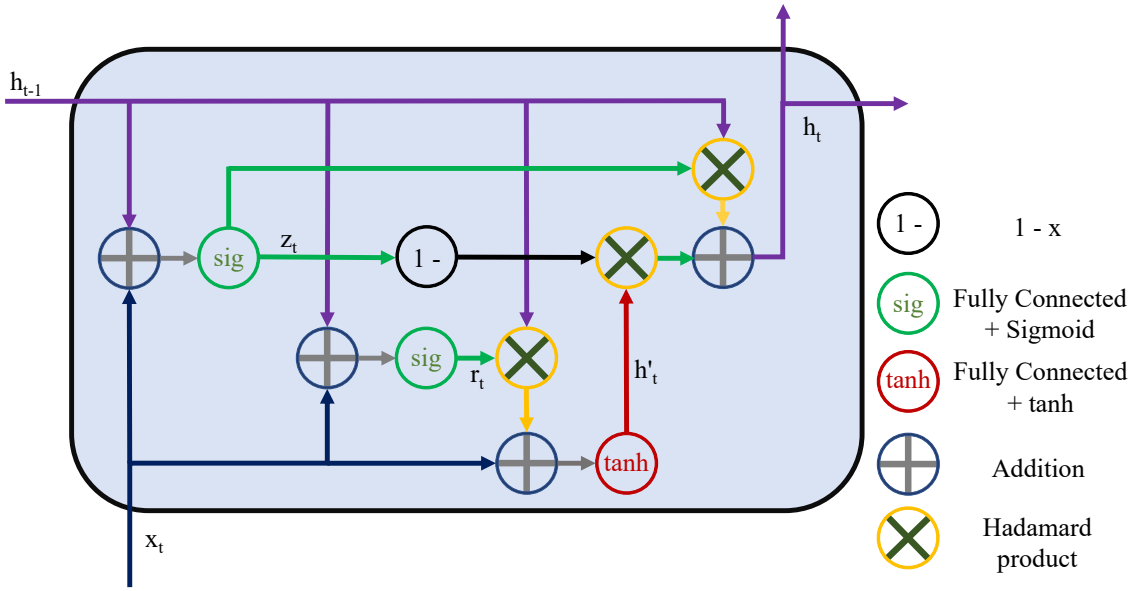

Figure 3.6: Schematic of a gated recurrent unit (GRU). An update gate z_t and a reset get r_t are used to control a proposed hidden state update h'_t to form the new hidden state h_t, which is also the output for a specific time step. Created using Powerpoint.

by randomly varying the stride between a value of one and two. This enables deeper ANNs with a larger number of abstraction levels, especially for small input sizes. Local average pooling layers work analogously to maximum pooling layers but propagate the mean value within the field of view. Global pooling operations follow the same principle, but result in a fixed output size. This is especially advantageous, if an ANN should be able to process differing input sizes, or if the classifier needs to be disentangled from the location of the object of interest within the data.

Recurrent Layers

Another way of processing structured data, typically applied for time-series and natural language data, is by employing recurrent cells. Contrary to feed-forward ANNs like the multi-layer perceptron (MLP) an recurrent neural network (RNN) has an "internal memory" or feedback loop, where the output of the previous input also influences the output of the current step. With this the context of a sequence can be modeled better. Furthermore, this concept allows for analysis of sequences of arbitrary lengths.

One of the simplest recurrent cells is the Elman cell, which employs a hidden state which is updated after each input and influences the whole latter sequence. A main problem of the Elman cell is that it is not potent enough to robustly capture long-term dependencies. One extension of the Elman cell is the so-called long short term

		$\hat{y}$ 1	$\hat{y}$ 0
y	1	TP	FN
	0	FP	TN

Table 3.1: A confusion matrix for a binary classification problem. A prediction $\hat{y}$ is compared to the ground truth y. Then the number of samples belonging to the respective true positive (TP), false negative (FN), false positive (FP), and true negative (TN) category are aggregated.

memory (LSTM) cell, which introduces an input forget and update gate to model both a hidden state and an additional cell state. An alternative being the GRU builds up on a similar concept but with a more efficient use of weights. A schematic of the inner workings of a GRU is depicted in Figure 3.6. Usually, GRUs are easier to train with similar performance as LSTM cells. As an alternative to specialized cells, 1D convolutions with varying dilation degrees can be used to analyze different frequencies of sequential data.

3.4 Evaluation

An important aspect to consider when developing machine learning approaches is to measure the performance of the algorithm with metrics of choice. Usually for classification tasks, these are derived from the confusion matrix as depicted in Table 3.1.

By comparing the ground truth label y to the prediction $\hat{y}$ each sample gets categorized as TP, FN, FP, or TN. The summed values over the whole test dataset are then used to calculate different metrics. A popular metric, which is easily understood is the accuracy (ACC) defined as

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.13}$$

being the ratio of the overall correctly classified samples. Other important metrics, especially in the medical context include the Sensitivity, Specificity and F1-score defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.14}$$

$$\text{Specifity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.15}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{3.16}$$

Here, Sensitivity denotes how many positive samples can be detected for the whole population of positive cases, while Specificity is a measure for how well negative

cases can be detected for the whole population of negative cases. The F1 score, on the other hand, measures how well a classifier works for both classes. However, it may be sensitive to class imbalance. Nevertheless, a method often used to report for problems suffering from class imbalance is the Matthew's correlation coefficient (MCC) defined as

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.17}$$

So far, only the setting where samples are assigned to a final class prediction was assessed. However, as ML algorithms are often capable of outputting a probability instead of a fixed prediction, threshold agnostic metrics are also of high interest. The most frequently reported metric for this is the area under the receiver operating characteristic curve (AUC). By calculating the sensitivity and the false positive rate (defined as $1 -$ specificity) for differing cut-off points the receiver operating characteristic curve can be plotted. By calculating the area under this curve, one can assess the classifiers capability to increase the TP rate while still maintaining a low number of FP samples.

CHAPTER 4

Related Work

With the theoretical background established, this chapter will offer an overview of methods pertaining to the research conducted within the scope of this thesis. The primary emphasis is placed on the classification of coronary artery disease (CAD) derived from coronary CT angiography (CCTA) scans. It should be noted that individual papers in the subsequent Chapters include sections covering related work, but due to the rapidly evolving nature of this research field, these sections are somewhat outdated. Therefore, this Chapter aims to present an improved overview of the current state of research at the time of writing this thesis. This Chapter does not include work related to the research described in Chapter 7, as that Chapter encompasses very recent publications. Structure-wise, this Chapter will first depict the clinical workflow for coronary plaque analysis, which is used as an analogy for the subsequent method designs. Next, an overview over the variety of input data representations is given, followed by discussions on different architectural choices. Finally, the targets of the respective related research are briefly described.

4.1 Clinical Background

To gain a deeper understanding of the methodologies employed by researchers for the classification of CAD from CCTA, it is essential to examine how this problem is typically addressed in clinical settings. In such scenarios, CAD grading typically follows a multi-stage pipeline. Initially, an optional step of localizing and segmenting of the heart is performed to assist downstream algorithms that can benefit from this supplementary information. Subsequently, the coronary centerlines are typically extracted using semi-automated or automated approaches, aiming to localize and allow reformatting of the anatomical region of interest. It is noteworthy that this

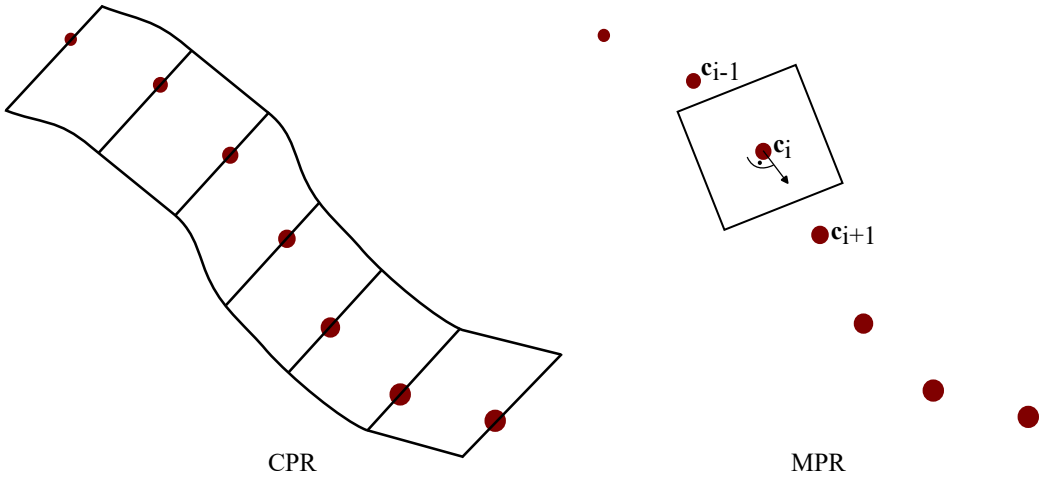

Figure 4.1: Commonly used vessel reformation strategies: the CPR view is sampled by interpolating a line for each centerline point, to gain a CPR plane. On the other hand, an MPR view is constructed by interpolating a plane orthogonal to the centerline direction at a specific centerline point $\mathbf{c}_i$. The direction is usually determined through finite differences of the neighboring points $\mathbf{c}_{i+1} - \mathbf{c}_{i-1}$. Created using Inkscape v1.22

step can be prone to errors and may necessitate manual correction by physicians, particularly in areas with significant calcifications or total occlusions. As an optional intermediate step, the vessel segments can be subdivided according to the american heart association (AHA) guidelines to facilitate downstream reporting of the location of found lesions. Following that, two types of reformatted vessel images are commonly obtained from the CCTA volume: the curved planar reformation (CPR) view and the multi planar reformation (MPR) view. The CPR view is generated by mapping all centerline points onto a curved plane, thereby preserving the vessel's approximate curvature and providing a longitudinal cross-sectional view. Conversely, MPR images are produced by interpolating a view for a single centerline point orthogonal to the direction of the centerline. These reformation strategies are depicted in Figure 4.1. Subsequently, the assumed healthy vessel diameter is measured at both proximal and distal reference positions, ideally free from any narrowing, and then compared to the most severe narrowing observed within the lesion. The CPR view often serves as a navigational aid, while the MPR view facilitates precise measurements.

4.2 Data Representation

Following this workflow, most related works assume a given centerline extraction algorithm or manually annotated centerlines. A majority of studies focus on the MPR

views, often stacking them to form overlapping cubes [Zrei 18, Zrei 19, Ma 21, Lin 22] or stacked to large volumes for whole branches [Cand 20, Gupt 20, Whit 21, Yang 21, Zhan 22, Chen 22, Pens 23, Gerb 23]. However, also encoding of single views for downstream analysis is an alternative [Pabl 19, Hamp 22]. This representation is well-motivated by the clinical workflow and reduces the algorithm's search space to the relevant image sections. Moreover, the crucial information for classification lies in the radial information of the vessel diameter at each centerline point along the centerline course, which can be effectively captured using this data representation. To delve deeper, Gupta et al. [Gupt 20] and White et al. [Whit 21] employ a cylindrical mapping technique, projecting images at every 10 degrees around the centerline to a plane, and organize the resulting 18 images into a 2×9 grid to encompass the entire vessel. This approach incorporates the knowledge of the cylindrical nature of vessels within the MPR volume format. Furthermore, the orthogonal interpolation of the MPR slices ensures a uniform spacing across multiple datasets, enhancing the generalizability of the approaches to datasets acquired with differing spacing, although fully closing this gap remains challenging. However, relying on prior centerline extraction may prove suboptimal in some cases, as the overall performance of the method is always limited by the quality of the centerline extraction. Another strategy involves using the CPR views as inputs [He 22, Paul 22]. While this representation preserves the vessel's curvature to some extent, it includes a significant amount of non-essential information, and accessing the relevant information is not as straightforward as in MPR views. A third category of approaches assesses the coronary arteries directly from the CCTA data without prior reformatting. These methods either still incorporate the centerline information by extracting 3D cubes at each centerline point [Viti 22] or, in the work of Jin et al. [Jin 22], leverage the whole CCTA volume as the input. However, in the latter case, a prior localization network is employed as a substitute to the centerline extraction to enable downstream focus on the relevant regions.

4.3 Deep Learning Architectures

When it comes to the employed architectures, there is a variety of approaches. One common method for analyzing the coronary arteries is to extract local features at specific centerline positions and then examine the sequence of features along the centerline. This approach aligns well with the clinical workflow, where the vessel is analyzed at different positions to obtain a lesion-wise assessment. The first

proposed deep learning-based approach utilizes a straightforward combination of a convolutional neural network (CNN) combined with bilateral gated recurrent units (GRUs) [Zrei 18]. However, later on the GRUs are replaced with the Transformer architecture [Ma 21] which is a logical extension as Transformers often outperform the traditional recurrent neural network (RNN) cells. Similarly, Viti et al. [Viti 22] extract local features using a 3D CNN. However, this is followed by a graph network to directly aggregate and analyze the global context.

Some approaches that take MPR slices as an input also incorporate components initially introduced for time-series analysis. For instance, Lin et al. [Lin 22] combine CNNs and long short term memory (LSTM) cells to segment the lumen area for downstream stenosis grading. Other approaches with this kind of input map the input to either a feature representation [Pabl 19] or alternatively estimate the vessel radius and plaque extend [Hamp 22]. These intermediate representations are then used as input for either classical methods based on Gaussian mixture models (GMMs) and support vector machines (SVMs)[Pabl 19] or a subsequent analysis with CNNs [Hamp 22]. The advantages and disadvantages of this data representation mostly align with those of using MPR views as input. However, these approaches are required to extract the 3D information solely from the downstream analysis. To overcome this limitation, Tejero-de-Pablos et al. [Pabl 19] address this issue by also leveraging longitudinal slices along the centerline direction.

Moving on to methods that take larger MPR volumes as an input, previous approaches employ 3D CNN architectures [Cand 20, Yang 21, Chen 22]. However, more recent architectures also embrace the Transformer architecture in the form of the Conv-Mixer architecture [Pens 23] or vision Transformers [Gerb 23]. Approaching the problem from a different perspective, Zhang et al. [Zhan 22] utilize a Mask region-based convolutional neural networks (R-CNN) for lesion detection and perform downstream classification on these regions. In general, 3D CNNs applied to classification tasks often struggle due to the high dimensionality of the input, which results in a singular output guiding the training of a large number of weights. A limitation of most of the mentioned approaches is that the clinically relevant information, exhibited radially to the centerline in the form of the lumen, is not easily assessable in Cartesian coordinates. This is because CNNs are inherently not rotationally invariant, which may cause varying predictions for inputs rotated around the centerline. Moreover, vessels in the MPR view typically have a cylindrical shape. Two aforementioned data representations encode the 3D information in 2D views exploiting a cylindrical mapping [Gupt 20, Whit 21]. As a benefit, these can leverage 2D CNNs for analysis, for which a larger number of pre-trained models exist, and they are generally better

explored for classification tasks. With a similar reasoning Paul et al. [Paul 22] and He et al. [He 22] employ CPR views to use 2D CNN architectures. As an outlier the work of Jin et al. [Jin 22] use a Mask R-CNN for both coronary artery as well as plaque candidate segmentation. The resulting segmentation masks and CCTA volume are then used to calculate Radiomics for the region, which are subsequently classified using gradient boosting decision trees.

4.4 Targets

An important aspect that has deliberately been withheld until now is the choice of the actual target for the various approaches discussed. The reason for withholding this information was that the fundamental task remains largely the same across the presented related works: analyzing the coronary arteries in relation to coronary artery disease. To provide a comprehensive overview, some of the related works aim to predict only the presence or absence of coronary artery disease [Gupt 20, Whit 21], while others focus on predicting the significance of stenosis, indicating the need for further assessment [Zrei 18, Pabl 19, Cand 20, Yang 21, Viti 22, He 22]. Additionally, some works aim to automatically assess the haemodynamic significance of stenosis [Hamp 22] or automatically assess the plaque morphology [Zrei 18, Yang 21, Chen 22, Zhan 22]. The remaining works attempt to directly derive the coronary artery disease-reporting data system (CAD-RADS) grade as well [Jin 22, Paul 22, Lin 22, Pens 23, Gerb 23], sometimes binned with respect to the clinically most relevant thresholds. However, it is important to note that all the presented methods are also capable of deriving a patient-wise score, and the limitation on other targets is typically driven by the availability of annotated data. As a side note, it is worth mentioning that all the works follow a detection process followed by lesion-wise grading, which is then propagated to the patient level for the most severe lesion. While this grading scheme is explainable, it does not incorporate a global context into the final grade and is sensitive to outliers.

PART II

Contributions

CHAPTER 5

Coronary Artery Plaque Characterization

5.1 Introduction

The research presented in this chapter is closely aligned with the related work discussed in Chapter 4. Since there are notable similarities in terms of the tasks and data utilized in both studies presented in this Chapter, this Section serves as a combined introduction to the topic. The two primary tasks addressed in these studies are as follows: first, classifying the presence of significant stenosis in previously annotated lesions with defined start and end points, and secondly, examining whether the lesion leads to a downstream revascularization decision. The selection of these objectives was motivated by the availability of lesion-wise stenosis degree and patient-wise revascularization decision annotations.

In the studies, a relatively small population of 95 patients is included, encompassing a total of 345 lesions. It should be noted that not all smaller lesions are necessarily included, as the labeling task did not involve an exhaustive annotation of all lesions. Consequently, the focus was solely on classifying lesions with defined start and end points, in contrast to related work, which typically includes lesion detection as well. While this simplification offers certain advantages, it also presents some limitations. Firstly, the consideration of lesions with varying lengths becomes crucial. Secondly, the amount of data provided to the network is limited to the small number of annotated regions, rather than the entire coronary artery trees. This, however, also shifts the data balance towards the more relevant regions. As a side-note: for

details on the used segmentation algorithms for annotation and mask creation, the reader is pointed towards our work describing the clinical research prototype used:

[Denz 21a] F. Denzinger, M. Wels, C. Hopfgartner, J. Lu, M. Schöbinger, A. Maier, and M. Sühling. "Coronary Plaque Analysis for CT Angiography Clinical Research". In: *Bildverarbeitung für die Medizin 2021*, pp. 223–228, Springer, 2021

5.2 Coronary Artery Plaque Characterization using Deep Learning and Radiomics

The first publication to be presented is:

[Denz 19] F. Denzinger, M. Wels, N. Ravikumar, K. Breininger, A. Reidelshöfer, J. Eckert, M. Sühling, A. Schmermund, and A. Maier. "Coronary artery plaque characterization from CCTA scans using deep learning and radiomics". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 593–601, Springer, 2019

The main contributions are:

- Prediction of the final revascularization decision from a lesion annotation using deep learning (DL) as the first reported approach.
- Evaluation of a Radiomics-based approach for coronary artery disease (CAD) assessment.
- Proposal of an extension to prior work by leveraging polar transform, fractional max pooling and a problem- and representation-specific network design.
- Introduction of a combination of Radiomic features and recurrent neural network (RNN)s for the task at hand.

5.2.1 Publication Overview

Building upon the previous introduction, this work proposes three distinct approaches. Firstly, a Radiomics-based approach is employed, which utilizes the plaque segmentation as input. By extracting a large number of features from the segmentation mask, an eXtreme Gradient Boosting (XGBoost) classifier is employed for intrinsic feature selection and classification. As a second approach, we extend the work of Zreik et al. [Zrei 19] which leverages a convolutional neural network (CNN) to extract local

features from a series of overlapping cubes obtained from a multi planar reformation (MPR) volume. This sequence is then analyzed using an RNN. The approach of Zreik et al. is well-motivated due to the characteristic narrowing and widening of coronary artery along the centerline direction. Still we aim to enhance its performance.

One aspect of potential improvement lies in the data representation. To embrace the cylindrical structure of coronary arteries, we transform all single axial slices into polar coordinates. This transformation not only emphasizes the cylindrical nature but also enhances the rotational invariance of the downstream CNN. Additionally, we adapt the CNN architecture to align with the characteristics of coronary arteries. Our proposed network primarily employs slice-wise, i.e. applied on individual axial cross sections, convolutions instead of 3D convolutions, combining them solely along the centerline direction using a final $1 \times 1 \times 1$ convolution. This architectural adjustment compels the network to predominantly learn radial features, which are later combined.

As a third approach, we combine the first two by leveraging the lumen segmentation mask and dividing it into a sequence of masks, similar to the deep learning-based approach. From these masks, shape-based Radiomics features are extracted, and finally, the sequence of features along the centerline direction is analyzed with an RNN.

In terms of results, a mixed picture emerges for the two objectives. For predicting significant stenosis, the approach of Zreik et al.'s [Zrei 19] outperforms our deep learning-based approach, while the purely Radiomics-based approach performs comparably. However, the combined approach significantly outperforms both, achieving an area under the receiver operating characteristic curve (AUC) of 0.96 compared to the original 0.89 [Zrei 19]. It is worth stating again that both Radiomics-based approaches require a prior segmentation, which is a heavy drawback. Regarding the revascularization decision, the adaptations made to the data representation and CNN design prove advantageous as the AUC improves from 0.80 to 0.84 compared to Zreik et al.'s approach [Zrei 19]. Once again, the Radiomics-based approaches outperform the others in this regard.

Coronary Artery Plaque Characterization from CCTA Scans using Deep Learning and Radiomics

Felix Denzinger[1,2], Michael Wels[2], Nishant Ravikumar[1], Katharina Breininger[1], Anika Reidelshöfer[3], Joachim Eckert[4], Michael Sühling[2], Axel Schmermund[4], and Andreas Maier[1]

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2] Computed Tomography, Siemens Healthcare GmbH, Forchheim, Germany
[3] University Clinic Frankfurt, Frankfurt am Main, Germany
[4] Cardioangiological Centrum Bethanien, Frankfurt am Main, Germany

Abstract. Assessing coronary artery plaque segments in coronary CT angiography scans is an important task to improve patient management and clinical outcomes, as it can help to decide whether invasive investigation and treatment are necessary. In this work, we present three machine learning approaches capable of performing this task. The first approach is based on radiomics, where a plaque segmentation is used to calculate various shape-, intensity- and texture-based features under different image transformations. A second approach is based on deep learning and relies on centerline extraction as sole prerequisite. In the third approach, we fuse the deep learning approach with radiomic features. On our data the methods reached similar scores as simulated fractional flow reserve (FFR) measurements, which - in contrast to our methods - requires an exact segmentation of the whole coronary tree and often time-consuming manual interaction. In literature, the performance of simulated FFR reaches an AUC between 0.79-0.93 predicting an abnormal invasive FFR that demands revascularization. The radiomics approach achieves an AUC of 0.84, the deep learning approach 0.86 and the combined method 0.88 for predicting the revascularization decision directly. While all three proposed methods can be determined within seconds, the FFR simulation typically takes several minutes. Provided representative training data in sufficient quantities, we believe that the presented methods can be used to create systems for fully automatic non-invasive risk assessment for a variety of adverse cardiac events.

Keywords: Plaque Characterization · Computer Aided Diagnosis · Coronary CT Angiography · Radiomics · Deep Learning.

1 Introduction

Cardiovascular diseases (CVDs) have persisted to be the leading cause of death across all developed countries [10]. Most CVDs are related to atherosclerotic

plaques in the associated arteries [11]. Two types of high risk plaque segments exist: functionally significant plaques, which narrow the lumen and immediately lead to cardiac ischemia, and vulnerable plaques, which can rupture and cause thrombus formation and adverse coronary syndromes (ACS) such as stroke or cardiac infarction.

The reference standard measure to judge whether a plaque segment is functionally significant and the corresponding vessel needs to be revascularized is the fractional flow reserve (FFR) value. FFR is defined as the pressure after a lesion relative to the pressure before the lesion, and is measured invasively [3]. As interventional procedures involving the heart have the risk of inducing adverse cardiac events, a non-invasive assessment of the type of plaque for further patient selection is highly desirable. A non-invasive approach for this task is simulated FFR, which aims to simulate the FFR values from coronary computed tomography angiography (CCTA) data using a fluid dynamics approach [14], which requires an exact segmentation of the whole coronary tree and computational mesh generation [16]. Sufficient segmentation quality can often only be achieved with time-consuming manual interaction.

Previously, radiomics have been proposed to represent quantitative image information which is inherent in the data but hard to interpret for human readers [8]. They include multiple intensity-, texture-, shape- and transformation-based metrics extracted from a lesion segmentation and have been shown to be able to characterize coronary plaques [6]. More recently, deep learning has been investigated to detect lesions with a high stenosis degree and to categorize the calcification grade of coronary plaques using a recurrent convolutional neural network (RCNN) [18]. In their work, they first extract multi planar reformatted (MPR) slices orthogonally to each centerline point. Next, they cut the resulting image stack into multiple overlapping cubes from which features are extracted using a 3D convolutional neural network (CNN). Finally, classification is achieved using a sequence analysis network.

In this work, we propose a fully automatic method to directly predict the clinical decision of revascularization based on single plague segments. We investigate three machine-learning approaches for classification: radiomic feature analysis, deep learning and a combination of both. For the first variant, radiomic features are extracted from each vessel segment based on the vessel segmentation in the region of interest. Contrary to the approach in [6], we do not perform data mining since it neglects cross-feature correlations. Instead, we train a bagging classifier, namely the XGBoost algorithm [1], which automatically detects relevant features and uses all information from the features. For the deep learning approach, we extend the approach presented in [18] by improving the data representation using a transformation of the image stack into a cylindrical coordinate system which allows for a more effective training of the network and reduce the risk of overfitting by using 2D instead of 3D convolutions. Thirdly, we propose a novel combination of both aforementioned approaches. After extracting a sequence of cubes along the centerline, we calculate the radiomic features of each cube using a plaque segmentation mask extracted a priori. The resulting

sequence of radiomic features is then recombined with a multi-layer perceptron (MLP) and subsequently analyzed using a sequence analysis network based on gated recurrent units (GRUs). We evaluate all variants on CCTA scans of 95 patients with a total of 345 plague segments using ten-fold cross validation and compare our results with simulated FFR.

2 Data

The data collection contains CCTA scans of 95 patients taken within a time span of 2 years with the same system. The decision for revascularization or further invasive assessment was based on different clinical indications, e.g., functional tests including cardiac stress MRI or MIBI SPECT, and was made by trained cardiologists. In some cases, identification of culprit lesions was additionally based on ECG abnormalities if these indicated a bad perfusion of a specific part of the heart muscle. In total, the data contained 345 lesions, which were annotated by defining their start and end centerline point and segmenting their inner and outer vessel wall using a fully automatic approach [9]. For all data sets, automatic centerline extraction was performed as described in [17]. For each main branch of the coronaries a label indicated whether it was revascularized or not. To obtain reliable labels on the segment level, we propagated a positive revascularization decision only to the segment with the highest stenosis grade. In order to allow for an comparison with the results in [18], we additionally assessed for all segments whether the stenosis grade was below or above 50 %. With this procedure 85 (24.64 %) lesions were labeled as having a high stenosis grade and 93 (26.97 %) as requiring revascularization.

3 Methods

3.1 Radiomic-based Classification

As mentioned, a multitude of shape-, intensity- and texture-based features is extracted under different image transformations from the lesion segmentation as radiomics. A detailed description of all radiomic features can be found in [7]. The extracted feature vector has a high dimensionality. Therefore, direct classification is hard to achieve due to the curse of dimensionality. To overcome this we used the XGBoost classifier [1], which calculates its prediction based on an ensemble of decision trees while minimizing a loss function based on the total ensemble prediction. Since new leaves are added based on greedy search, only relevant non-redundant features are selected during training. Features were calculated using the open source PyRadiomics library [5] selecting all possible features under all transformations.

3.2 Deep Learning-based Classification

The second approach is based on deep learning and can be separated into several steps: data extraction, local feature extraction and sequence analysis. An

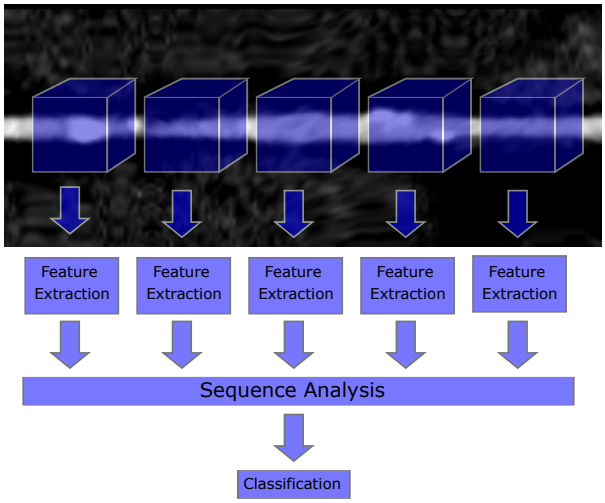

Fig. 1. Algorithm overview: extraction of a sequence of cubes along the centerline is followed up by a feature extraction method – either with a convolutional neural network or the PyRadiomics module. The resulting sequence of features is then analyzed by a sequence analysis neural network.

overview of the workflow is shown in Figure 1. First, MPR slices are extracted orthogonally to each point of the centerline in the segment. Then, the resulting image stack is cut into multiple overlapping cubes. The extracted cubes are transformed to polar coordinates to allow for a better representation for the neural network. The motivation behind this lies in the assumption that features that characterize lesions are formed radially to the centerline and vary along the centerline direction. The slices of each transformed cube are then used as input to a 2D-CNN that performs a slice-wise feature extraction. This is followed by 1x1 convolutions in centerline direction that recombine and fuse the information across a cube to perform a local feature extraction. The architecture of the feature extraction network is depicted in Figure 2, alongside the 3D-CNN network proposed in [18] that we evaluate for comparison. To obtain a final classification, we perform a sequence analysis using a two layer recurrent neural network (RNN) using gated recurrent units [2] on the features extracted from the cubes with the centerline direction as "temporal" dimension. Based on the assumption that information about the plaque composition is contained in both directions of the centerline, we perform the sequence analysis in a bidirectional fashion.

3.3 Combined Approach

A common way to train neural networks with a limited amount of data is to use pretrained models, which comprise relevant image features already learned

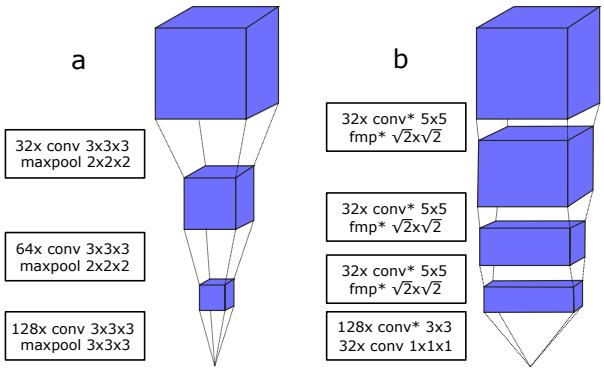

Fig. 2. Feature extraction model overview: a) model as described in [18]. b) our proposed model. * denotes a slice-wise operation, and fmp denotes fractional max pooling [4], which allows a pooling size smaller than 2 which enables feature extraction from intermediate scales.

on different data sets. However, this is difficult when dealing with medical data, since data of different organs, modalities and use cases are often not correlated and three-dimensional. To overcome this, non-deep learning feature extraction methods can be used and combined with deep learning. Therefore, we combine the above mentioned radiomic and deep learning approaches. Again, the vessel was sliced in a sequence of overlapping volume cubes, but now the feature extraction was performed using the PyRadiomics library and the vessel segmentation of the plaque segment. Since preliminary experiments suggested the shape-based feature group to be the most important for estimating both the revascularization decision and the stenosis degree, we focused on these features. The resulting sequence of radiomic feature vectors was further evaluated using a three layer MLP before analyzing the sequence with bidirectional GRUs.

4 Experiments and Results

We evaluate the proposed approach for binary stenosis grade classification (high-degree stenosis $> 50\,\%$, low-degree stenosis $< 50\,\%$) to allow for a direct comparison with [18] and for the prediction of clinical revascularization decisions. For all experiments, evaluations were performed using ten-fold cross validation with patient-wise stratified splitting. For the neural network based methods, $20\,\%$ of the training data was set aside as validation set in each fold. For each fold, the networks were trained for 50 epochs and the model that performed best on the validation set was selected for evaluation on the test set. For the CNN-based methods, data augmentation was performed in form of random rotation, mirroring along the x-axis, translation and additive Gaussian noise, and we resampled the data during batch generation to achieve class balance. In order to normalize

Table 1. Evaluation results for stenosis degree prediction on lesion-level. The results in the first row are copied from [18].

Model/metric	AUC	Acc	F1	PPV	NPV	Sens	Spec	MCC
3D-RCNN [18] (orig data)	–	0.94	0.64	0.65	0.97	0.63	0.97	0.60
3D-RCNN [18] (our data)	0.89	0.85	0.67	0.58	**0.94**	0.79	0.86	0.59
2D-RCNN + polar transform	0.86	0.87	0.64	0.68	0.91	0.60	**0.93**	0.56
Radiomics + XGBoost	0.89	0.84	0.69	0.69	0.90	0.68	0.90	0.58
Radiomics + GRUs	**0.96**	**0.92**	**0.95**	**0.94**	0.82	**0.96**	0.75	**0.74**

our data, histogram equalization was performed for each approach before feature extraction. To evaluate our approaches, we computed the area under the receiver operating characteristic curve (AUC), accuracy (Acc), F1-score, positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity and the Matthews correlation coefficient (MCC).

4.1 Stenosis Grade Classification

The classification results of the stenosis grade classification for the proposed methods and the 3D-CNN approach proposed in [18] are shown in Table 1. Compared to the results reported in [18], the performance of the 3D-RCNN approach on our dataset is lower. The main reason for this is likely the size of the respective data set, which was much smaller in our case. The proposed 2D-RCNN and radiomics approach achieved results on par with the 3D-RCNN. However, our combined approach outperformed all three other methods by a large margin (AUC 0.96 vs. 0.89 for 3D-RCNN/Radiomics+XGBoost and 0.86 for 2D-RCNN).

4.2 Classification of Revascularization Decision

The metrics for the revascularization prediction can be seen in Table 2. Since there exists a lot of variance with respect to the reference standard simulated

Table 2. Evaluation results for revascularization decision prediction on lesion-level.

Model/metric	AUC	Acc	F1	PPV	NPV	Sens	Spec	MCC
Simulated FFR best [13][a]	0.93	0.86	–	0.61	0.95	0.84	0.86	–
Simulated FFR worst [12][a]	0.79	0.69	–	0.56	0.84	0.61	0.89	–
3D-RCNN [18] (our data)	0.80	0.76	0.55	0.45	**0.91**	0.72	0.77	0.42
2D-RCNN + polar transform	0.84	0.82	0.57	0.60	0.88	0.54	0.91	0.46
Radiomics + XGBoost	0.86	0.86	0.62	0.69	0.89	0.56	**0.94**	0.54
Radiomics + GRUs	**0.88**	**0.87**	**0.92**	**0.90**	0.74	**0.95**	0.61	**0.60**

[a] Simulated FFR is compared to abnormal invasive FFR instead of revascularization decision

FFR, we compare our approaches to the best [13] and worst [12] results reported in the review paper of [15]. Note that simulated FFR is compared to an abnormally high invasive FFR value rather than the revascularization decision in the referenced publications, with both targets being highly correlated. The experiments in [12, 13] were performed on different non-publicly available data sets. Comparing the two RCNN networks, our proposed method performed slightly better. This indicates that features other than the stenosis degree are relevant for the revascularization decision, and that transforming the image data into the polar space was beneficial. The radiomics approach outperformed both deep learning methods, while our combined approach again performed best.

5 Discussion and Conclusion

Identifying functionally significant stenosis in a non-invasive setup is an important task to improve clinical outcomes. We presented and compared three machine-learning methods for the prediction of stenosis degree and clinical revascularization decision based on CCTA scans: Radiomics combined with boosting trees, a convolutional recurrent neural network, and an approach that combines shape-based radiomics and recurrent neural networks. We were able to show that all methods were able to differentiate stenosis grade $> 50\,\%$ and $< 50\,\%$, and reliably identify plaque lesion that were later revascularized. Across both tasks, the combined approach performed best, also compared to results reported in literature. The combined approach comes at a cost of a higher computation time of up to 2 seconds compared to only milliseconds for the RCNN approaches and requires a prior segmentation of the vessel lumen in the region of the plague segment. Still, the additional computation time does not pose a clinical limitation and the lumen segmentation is easily obtainable in an automated fashion. In contrast to this, simulated FFR requires an exact segmentation of the whole coronary tree and computation times of several minutes. For classification of revascularization, we showed that the performance of the proposed methods lies well within the range of prediction performance obtained by FFR simulation in literature. Given data with appropriate annotations, we believe that our methods would also perform well in identifying so-called culprit lesions that cause adverse cardiac events. Interestingly, the performance difference between the combined approach and the RCNN methods leads to the conclusion that extracting the shape-based features is highly relevant for differentiating lesions, but is harder to achieve for a completely data driven CNN-based feature extractor and may require a larger training data set. If only limited data is available, the combined approach proposed here seems to be promising, as predefined features and data-driven learning are fused. A limitation of the current study is that no simulated FFR values for the data set under investigation were available, which will be subject of future work. Additionally, the results will be validated on additional data collections that also include the invasive FFR measurements for comparison.

Disclaimer The methods and information here are based on research and are not commercially available.

References

1. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 785–794. ACM (2016)
2. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint: 1406.1078 (2014)
3. Cury, R.C., et al.: Coronary artery disease-reporting and data system (CAD-RADS): an expert consensus document of SCCT, ACR and NASCI: endorsed by the ACC. JACC: Cardiovascular Imaging **9**(9), 1099–1113 (2016)
4. Graham, B.: Fractional max-pooling. arXiv preprint arXiv:1412.6071 (2014)
5. van Griethuysen, J., et al.: Computational radiomics system to decode the radiographic phenotype. Cancer Research **77**(21), e104–e107 (2017)
6. Kolossváry, M., et al.: Radiomic features are superior to conventional quantitative computed tomographic metrics to identify coronary plaques with napkin-ring signclinical perspective. Circulation: Cardiovascular Imaging **10**(12) (2017)
7. Kolossváry, M., et al.: Cardiac computed tomography radiomics: a comprehensive review on radiomic techniques. Journal of Thoracic Imaging **33**(1), 26–34 (2018)
8. Lambin, P., et al.: Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer **48**(4), 441–446 (2012)
9. Lugauer, F., Zheng, Y., Hornegger, J., Kelm, B.M.: Precise lumen segmentation in coronary computed tomography angiography. In: International MICCAI Workshop on Medical Computer Vision. pp. 137–147. Springer (2014)
10. Mendis, S., Davis, S., Norrving, B.: Organizational update: the World Health Organization global status report on noncommunicable diseases 2014. Stroke **46**(5), e121–e122 (2015)
11. Naghavi, M.: From vulnerable plaque to vulnerable patient: a call for new definitions and risk assessment strategies. Part II. Circulation **108**, 1772–1778 (2003)
12. Nakazato, R., et al.: Noninvasive fractional flow reserve derived from computed tomography angiography for coronary lesions of intermediate stenosis severity: results from the DeFACTO study. Circulation: Cardiovascular Imaging **6**(6), 881–889 (2013)
13. Nørgaard, B.L., et al.: Diagnostic performance of noninvasive fractional flow reserve derived from coronary computed tomography angiography in suspected coronary artery disease: the NXT trial (Analysis of Coronary Blood Flow Using CT Angiography: Next Steps). Journal of the American College of Cardiology **63**(12), 1145–1155 (2014)
14. Taylor, C.A., Fonte, T.A., Min, J.K.: Computational fluid dynamics applied to cardiac computed tomography for noninvasive quantification of fractional flow reserve: Scientific basis. Journal of the American College of Cardiology **61**(22), 2233–2241 (2013)
15. Tesche, C., et al.: Coronary CT angiography–derived fractional flow reserve. Radiology **285**(1), 17–33 (2017)
16. Wels, M., Lades, F., Hopfgartner, C., Schwemmer, C., Sühling, M.: Intuitive and accurate patient-specific coronary tree modeling from cardiac computed-tomography angiography. In: The 3rd interactive Medical Image Computing Workshop. pp. 86–93. Athens (2016)

17. Zheng, Y., Tek, H., Funka-Lea, G.: Robust and accurate coronary artery centerline extraction in CTA by combining model-driven and data-driven approaches. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 74–81. Springer, Nagoya (2013)
18. Zreik, M., et al.: Automatic detection and characterization of coronary artery plaque and stenosis using a recurrent convolutional neural network in coronary CT angiography. arXiv preprint:1804.04360 (2018)

5.3 Deep Learning Algorithms for Coronary Plaque Characterization

Delving deeper into the topic of coronary artery plaque characterization with the aim of improving previous results without the need of a segmentation, the following research was published:

[Denz 20b] F. Denzinger, M. Wels, K. Breininger, A. Reidelshöfer, J. Eckert, M. Sühling, A. Schmermund, and A. Maier. "Deep learning algorithms for coronary artery plaque characterisation from CCTA scans". In: *Bildverarbeitung für die Medizin 2020*, pp. 193–198, Springer, 2020

In this work we contributed to the field of research:

- An evaluation of length normalization strategies for lesions of differing sizes.
- Examining how test time augmentation (TTA) influences the performance of three different approaches for coronary plaque analysis.
- Proposal of a novel 2.5D DL-based approach for coronary plaque analysis.

5.3.1 Publication Overview

Once again, we extend upon the introduction of this Chapter. A crucial aspect addressed in the publication at hand is the handling of lesions with varying lengths. As mentioned earlier, lesions with defined start and end points were utilized as input for this research. However, to facilitate network training, input batches in a training run typically have the same shape. Although it is possible to aggregate gradients over multiple inputs to stabilize gradient descent, this approach is time-consuming, and the network may overfit to the lengths of individual lesions. Hence, this paper aims to evaluate several straightforward approaches to create samples of the same shape, including padding with zeros, interpolating to the longest lesion, and interpolating to an intermediate length.

Furthermore, we assess the impact of utilizing TTA on the performance of deep learning-based approaches. This involves predicting for different equally spaced rotation angles around the centerline and averaging the predictions. We consider the approaches of Zreik et al. [Zrei 19] and Tejero-de-Pablos et al. [Pabl 19], as well as an additional novel approach. The proposed approach involves using two orthogonal longitudinal slices from the MPR volume along the centerline direction as input. This allows for direct assessment of radial information along the centerline, with the

second orthogonal view ensuring the preservation of some form of 3D information. To prevent suboptimal angles for the two views, the application of TTA is crucial for this method. The advantages of this approach include reduced dimensionality of the input data, increased variability of the training data, and a simpler architecture.

Our results indicate that resizing to an intermediate size yields the best performance, and the application of TTA improves performance across all methods. Furthermore, our novel architecture performs as well as the more complex approach of Zreik et al. [Zrei 19] for both tasks on our data. Overall, we achieved an AUC of 0.92 for predicting the degree of significant stenosis and an AUC of 0.90 for predicting subsequent revascularization decisions.

Deep Learning Algorithms for Coronary Artery Plaque Characterisation from CCTA Scans

Felix Denzinger[1,2], Michael Wels[2], Katharina Breininger[1], Anika Reidelshöfer[3], Joachim Eckert[4], Michael Sühling[2], Axel Schmermund[4], Andreas Maier[1]

[1]Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany
[2]Computed Tomography, Siemens Healthineers, Forchheim, Germany
[3]University Clinic Frankfurt, Frankfurt am Main, Germany
[4]Cardioangiological Centrum Bethanien, Frankfurt am Main, Germany
felix.denzinger@fau.de

Abstract. Analysing coronary artery plaque segments with respect to their functional significance and therefore their influence to patient management in a non-invasive setup is an important subject of current research. In this work we compare and improve three deep learning algorithms for this task: A 3D recurrent convolutional neural network (RCNN), a 2D multi-view ensemble approach based on texture analysis, and a newly proposed 2.5D approach. Current state of the art methods utilising fluid dynamics based fractional flow reserve (FFR) simulation reach an AUC of up to 0.93 for the task of predicting an abnormal invasive FFR value. For the comparable task of predicting revascularisation decision, we are able to improve the performance in terms of AUC of both existing approaches with the proposed modifications, specifically from 0.80 to 0.90 for the 3D-RCNN, and from 0.85 to 0.90 for the multi-view texture-based ensemble. The newly proposed 2.5D approach achieves comparable results with an AUC of 0.90.

1 Introduction

Cardiovascular diseases (CVDs) remain the leading cause of natural death [1]. In diagnosis and treatment of CVDs, the identification of functionally significant atherosclerotic plaques that narrow the coronary vessels and cause malperfusion of the heart muscle plays an important role. In clinical practice, this is typically assessed using fractional flow reserve (FFR) measurements [2].

This measurement is performed minimally invasively and therefore induces a small but existing risk to the patient. A non-invasive modality capable of visualising and assessing coronary artery plaque segments is coronary computed tomography angiography (CCTA). Current research tries to simulate the FFR value from CCTA scans [3]. Approaches based on this mostly rely on a prior segmentation of the whole coronary tree which is computationally intensive, prone to errors and may need manual corrections [4].

In this work, we investigate three lumen-extraction independent deep learning algorithms for the task of predicting the revascularisation decision and the

significance of a stenosis on a lesion-level. We propose a multi-view 2.5D approach, which we compare with two previously published methods, a 3D-RCNN approach [5] and a multi-view texture-based ensemble approach [6]. Additionally, we introduce adaptions to improve the performance of all approaches on our task. These include resizing lesions to an intermediate length instead of padding them and the usage of test-time augmentations. Also, we propose to use a different feature extraction backbone than described in [6] for the respective approach. Note that both reference approaches were originally used to detect lesions and characterise them. Contrary to this we characterise annotated lesions with a defined start and end point.

2 Material and Methods

2.1 Data

The data collection used contains CCTA scans from 95 patients with suspected coronary artery disease taken within 2 years at the same clinical site. For each patient, the resulting clinical decision regarding revascularisation was made by trained cardiologists, based on different clinical indications. This decision was monitored on a branch level. Lesions were annotated using their start and end point on the centerline, which was extracted automatically using the method described in [7]. We binarise the stenosis grade, which is estimated based on the lumen segmentation and defined as the ratio between the actual lumen and an estimated healthy lumen, using a threshold of 50 % according to [2]. The branch-wise revascularisation decision is propagated only to the lesion with the highest stenosis grade in branches known to be revascularised. Of the total of 345 lesions in our data set, 85 lesions exhibit a significant stenosis grade, and 93 require revascularisation.

2.2 Methods

3D-RCNN The first network we use is identical to the method described in [5]. In this approach, after extracting the coronary centerlines, a multi-planar reformatted (MPR) image stack is created by interpolating an orthogonal plane for each centerline point. Next, the MPR image stack is cut into a sequence of 25 overlapping cubes with size 25x25x25 and a stride of 5. During training, data augmentation using random rotations around the centerline and random shifts in all directions is used. Moreover, the data set is resampled for batch creation to achieve class balance during training. Since detection instead of sole characterisation is performed in [5], padding the inputs to the same length was not needed in their work.

Texture-based Multi-view 2D-CNN The second baseline approach is described in reference [6]. A VGG-M network backbone pretrained on the ImageNet challenge dataset is used as a texture-based feature extractor. The extracted

features are encoded as Fisher vectors and used for classification using a linear support vector machine. As inputs for this classification pipeline, different 2D views of the MPR image stack are combined for a final vote.

2.5D-CNN Both aforementioned methods utilize a sliced 3D representation of the lesion or a multitude of 2D representations, which is computationally expensive to obtain and to process by the subsequent machine learning pipeline. To mitigate this, we propose a 2.5D multi-view approach as shown in Figure 1. From the MPR image stack, only two orthogonal slices are selected, concatenated and forwarded to a 2D-CNN.

Modifications In this work, we examine the effect of three different padding strategies for all three approaches: zero-padding, stretching the volume stack to the longest lesion and resizing all lesions to an intermediate size. Stretching and squeezing of the image stacks along the centerline is performed with linear interpolation. Each MPR image stack for each lesion has a resolution of 64x32x32 and 170x32x32 after padding depending on the method used. For the 3D-RCNN approach, we downscale the y and x dimension further to 25x25 to match the original algorithm described in [5]. For augmentation of the data set all single volumes are rotated around the centerline in steps of 20°, which leads to an 18 times larger data collection. In order to create valid rotational augmentations of the image stack without cropping artefacts, we cut out a cylindrical ROI and set all values around it to zero. We confirmed in preliminary experiments that this computationally cheaper procedure does not to impact the results compared to cutting out a rotated view from the original data. In contrast to [5,6], no class resampling was necessary during training, since the class imbalance is not as severe for classification given the start and end point of a lesion compared to detecting lesions as well. Instead of the originally proposed VGG-M backbone used in [6], we use the VGG-16 network architecture as a backbone since it was already shown to yield better performance in the original paper on texture-based filter banks [8]. The data set was normalised to fit ImageNet statistics. We also evaluate the performance of this approach using a pretrained Resnet50 architecture [9] as backbone.

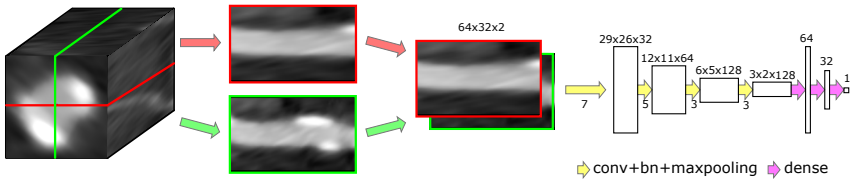

Fig. 1. Algorithm overview: Extraction of two orthogonal views of the lesion of interest. These are concatenated and then used as an input for a 2D-CNN (conv = convolutional layer, bn = batch normalisation layer, dense = fully connected layer).

Table 1. Results for predicting stenosis degree prediction on a lesion-level (18 and 8 correspond to the amount of views considered for data augmentation during training and test time, + = single view classification, ∗ = combined view classification, | | = resizing to intermediate size, | → | = resizing to the longest sequence).

Model/Metric	AUC	Accuracy	F1-score	Sensitivity	Specificity	MCC
3D-RCNN[5][10]	0.89	0.85	0.67	**0.79**	0.86	0.59
3D-RCNN[5]$^{18\|\ \|}$	**0.92±0.03**	0.88±0.02	0.69±0.06	0.68±0.11	0.93±0.03	0.62±0.06
2D[6]$^{8*\|\rightarrow\|VGG}$	0.85±0.07	0.86±0.04	0.62±0.10	0.56±0.15	0.94±0.02	0.54±0.12
2D[6]$^{18+\|\ \|RES}$	0.78±0.04	0.82±0.03	0.61±0.05	0.70±0.08	0.85±0.03	0.50±0.06
2D[6]$^{18*\|\ \|RES}$	0.90±0.04	0.87±0.03	0.68±0.08	0.71±0.13	0.91±0.03	0.60±0.09
2.5D$^{18+\|\ \|}$	**0.92±0.03**	0.89±0.02	0.70±0.06	0.64±0.10	0.95±0.03	0.64±0.06
2.5D$^{18*\|\ \|}$	**0.92±0.03**	**0.90±0.02**	**0.71±0.07**	0.64±0.10	**0.96±0.03**	**0.66±0.08**

Evaluation No hyperparameter optimisation is performed. Parameters are either taken from the references or default values are used. To reduce the influence of random weight initialisation and other random effects on the results, we repeat a 5-fold cross validation with five different initialisations, leaving a total of 25 splits. All splits are performed patient-wise. We also use the aforementioned rotational augmentation during test-time, and compare how the mean prediction over all rotations performs in comparison to a single input.

3 Results

The most important results are provided in Table 1, Table 2 and Figure 2. The results for the 3D-RCNN approach are also compared to the results of our

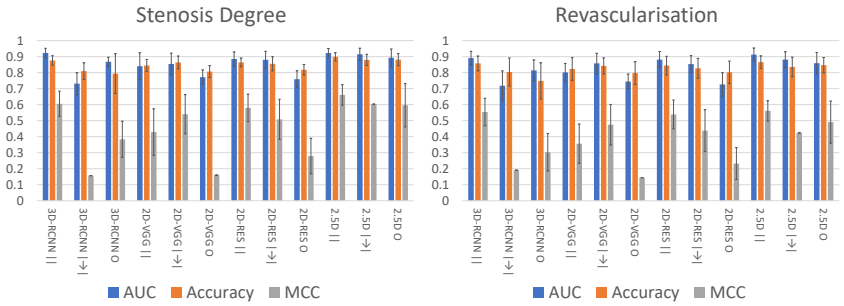

Fig. 2. Mean performance and standard deviation of all approaches for different padding strategies. These experiments are performed using only 8 instead of the 18 views (| | = resizing to intermediate size, | → | = resizing to the longest sequence, O = zero padding or no padding for the texture-based approach, MCC = Matthews correlation coefficient).

Table 2. Results for predicting the revascularisation decision on a lesion-level (Abbreviations as in Table 1).

Model/Metric	AUC	Accuracy	F1-score	Sensitivity	Specificity	MCC
3D-RCNN[5][10]	0.80	0.76	0.55	**0.72**	0.77	0.42
3D-RCNN[5]$^{18\vert\ \vert}$	0.90±0.05	0.84±0.10	0.63±0.10	0.65±0.13	0.90±0.12	0.53±0.11
2D[6]$^{8*\vert\rightarrow\vert VGG}$	0.86±0.06	0.84±0.05	0.56±0.12	0.49±0.16	0.93±0.02	0.47±0.13
2D[6]$^{18+\vert\ \vert RES}$	0.77±0.06	0.81±0.03	0.60±0.06	0.68±0.12	0.84±0.02	0.48±0.07
2D[6]$^{18*\vert\ \vert RES}$	0.90±0.06	0.85±0.05	0.66±0.07	0.70±0.16	0.89±0.04	0.57±0.10
2.5D$^{18+\vert\ \vert}$	**0.90±0.04**	0.87±0.05	0.65±0.05	0.60±0.11	0.94±0.04	0.58±0.06
2.5D$^{18*\vert\ \vert}$	**0.90±0.04**	**0.88±0.05**	**0.67±0.06**	0.61±0.11	**0.95±0.04**	**0.60±0.07**

previous work [10], where similar experiments are performed on the same data set as here but with the workflow described in [5], zero-padding and a different cross validation strategy. From the three padding methods examined, resizing all volume stacks of the data collection to one intermediate size yields the best results for most network approaches except for the texture-based approach with the VGG-16 backbone, where resizing all lesions to the size of the largest volume performs best. Interestingly, the same algorithm workflow with the Resnet50 backbone performs differently in that regard. A hypothesis that can be drawn from the intermediate padding performing best is that this scale provides on the one hand roughly the same amount of information per sample while on the other hand also keeping the input size in a range where it can be processed better. For the 3D-RCNN, we only look at classification in this work, in contrast to the task in [5] which included the detection of lesions. For this target, the proposed adaptations to the workflow in terms of padding strategy and not resampling the data set during batch creation improves the performance of both predicting the stenosis degree and the revascularisation decision from an AUC of 0.89 to 0.92, and 0.80 to 0.90, respectively. Having a more powerful feature extractor network for the texture-based approach combined with slightly more data augmentation improves the AUC by 0.05 for classifying stenosis significance, and by 0.04 for classifying revascularisation decision. The method performs considerably better when using test augmentations than without. Our proposed approach performs similar to the other two approaches, outperforming them by a small margin with an AUC of 0.92/0.90 for predicting a significant stenosis/revascularisation decision. Interestingly, test augmentations only yield a small improvement. This suggests that the method already has all necessary information to predict the task at hand from two orthogonal slices.

4 Discussion

In this paper, we compared and improved three segmentation independent deep learning-based algorithms for predicting both significant stenosis degree and clinical revascularisation decision for lesions annotated with a start and end point.

We obtained comparable results for each method. Our proposed method – a 2.5D approach – slightly outperforms the other approaches and requires fewer views compared to the method previously described in [6]. Therefore, a faster training procedure and inference is possible. In future work, we will examine whether this method is also capable of detecting lesions instead of just classifying them, and whether it is able to predict an abnormal FFR value.

Disclaimer The methods and information here are based on research and are not commercially available.

References

1. Mendis S, Davis S, Norrving B. Organizational update: the World Health Organization global status report on noncommunicable diseases 2014. Stroke. 2015;46(5):e121–e122.
2. Cury RC, et al. Coronary artery disease-reporting and data system (CAD-RADS): an expert consensus document of SCCT, ACR and NASCI: endorsed by the ACC. JACC: CI. 2016;9(9):1099–1113.
3. Taylor CA, Fonte TA, Min JK. Computational Fluid Dynamics Applied to Cardiac Computed Tomography for Noninvasive Quantification of Fractional Flow Reserve: Scientific Basis. JACC. 2013;61(22):2233–2241.
4. Wels M, Lades F, Hopfgartner C, Schwemmer C, Sühling M. Intuitive and Accurate Patient-Specific Coronary Tree Modeling from Cardiac Computed-Tomography Angiography. In: The 3rd interactive MIC Workshop; 2016. p. 86–93.
5. Zreik M, et al. A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography. IEEE Transactions on Medical Imaging. 2018;38(7):1588–1598.
6. Tejero-de Pablos A, et al. Texture-Based Classification of Significant Stenosis in CCTA Multi-view Images of Coronary Arteries. In: MICCAI. Springer; 2019. p. 732–740.
7. Zheng Y, Tek H, Funka-Lea G. Robust and accurate coronary artery centerline extraction in CTA by combining model-driven and data-driven approaches. In: MICCAI. Springer; 2013. p. 74–81.
8. Cimpoi M, Maji S, Vedaldi A. Deep filter banks for texture recognition and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3828–3836.
9. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.
10. Denzinger F, et al. Coronary Artery Plaque Characterization from CCTA Scans Using Deep Learning and Radiomics. In: MICCAI. Springer; 2019. p. 593–601.

5.4 Discussion

Within this chapter a variety of approaches to predict coronary artery disease, specifically focusing on significant stenosis degree and the revascularization decision, were presented. Our research has also addressed the challenge of handling lesions with varying lengths and highlighted the significant influence of TTA on model performance for these tasks. However, it is important to discuss how our approaches compare to more recently proposed methods and the general advantages and disadvantages of our techniques.

Regarding related work, Ma et al. [Ma 21], who also build upon Zreik et al.'s approach [Zrei 19], focus on improving the time-series analysis component. Interestingly, their experiments demonstrated that our network architecture, proposed in our initial work [Denz 19], outperformed Zreik et al.'s method on their dataset on the task of significant stenosis classification while we reported opposite behaviour on this task. This finding not only highlights that our proposed modifications to the original method, also work for the task of lesion detection, but also sheds light on a recurring issue in this research field: the lack of publicly available data collections for direct comparisons and the limited amount of data utilized in most studies. As a result, without direct comparisons, it is challenging to draw conclusive assessments regarding the best performing overall methods.

Nevertheless, our proposed approaches aim to enhance data representation and streamline network training for the targeted tasks. This is achieved by transforming individual axial slices into polar coordinates or by directly utilizing longitudinal slices. We also strive for simplicity in our architectures while incorporating task-specific knowledge. However, it is important to note that our best performing method from our previous publication requires a prior segmentation, which introduces an additional prerequisite. However, this segmentation step can be automated and integrated into the grading system, as demonstrated by Lin et al. [Lin 22]. The integration of segmentation improves the interpretability of the entire system, reducing its black box nature inherent to our approaches, but also introduces an extended pipeline that may be prone to error propagation.

Therefore, it might be worthwhile to explore the possibility of combining a black box approach, which can incorporate qualitative aspects, with segmentation-based approaches that focus more on quantitative aspects. This combination could potentially leverage the advantages of both approaches. Furthermore, it would be valuable to investigate whether disagreements between these different directions can be utilized to perform learning with abstention, thereby improving the overall performance.

In terms of comparing our approaches with methods that work directly on larger branches instead of cubes, there is currently no work that has performed a direct comparison. Theoretical advantages of our work include the sparser data representation, which still captures the relevant information, thereby potentially improving efficiency and reducing computational requirements.

In conclusion, our research has contributed by proposing various approaches to predict CAD. While we have achieved promising results, further studies with larger datasets and direct comparisons to recent approaches are needed to determine the best performing methods in this field. Additionally, the trade-offs between simplicity and performance, as well as the integration of prior segmentation, require careful consideration in the development of future diagnostic systems for CAD. Furthermore, exploring the combination of different approaches and investigating the potential of learning with abstention could lead to advancements in this domain.

CHAPTER 6

Coronary Artery Disease Classification

6.1 Introduction

Again, the related work presented in Chapter 4 is highly relevant to the research presented in this chapter. Additionally, there are significant similarities between the two subsequent publications, making it appropriate to provide a joint introduction. In both studies, the primary objective is the automated scoring of the coronary artery disease-reporting data system (CAD-RADS) grade, as described in Section 1.1.2, using deep learning (DL). Moreover, similar to the research presented in Chapter 5, the design of the methods is influenced by the availability of annotations for the given data collection.

The dataset used for both studies comprises approximately 2,900 patients, with a minor discrepancy in the number of patients between the two publications due to the automated pre-processing pipeline encountering issues for a limited number of cases. The annotations for this dataset include the patient-wise CAD-RADS score, Agatston score (AS) score, and stenosis grade labels for coronary subsegments labed according to the american heart association (AHA) guidelines. However, the dataset lacks centerlines labeled with respect to their AHA segments. Consequently, the stenosis-grade labels are weak labels, as their accuracy depends on the accuracy of the employed automated centerline labeling step and how well it matched the individual human annotator. Furthermore, the exact location of a lesion within a segment is unknown, and there may be instances of multiple lesions or lesions spanning across

segment boundaries. Therefore, the research conducted here primarily focuses on predicting the patient-wise CAD-RADS score, as it provides the most reliable labels. The AS score and stenosis degree labels are used as auxiliary targets. Addressing these challenges, the key tasks involve finding an appropriate data representation and designing an architecture to effectively aggregate information from the entire coronary tree for patient-wise assessment. Similar to Chapter 5, another challenge is handling segments of varying lengths.

6.2 Automatic CAD-RADS Scoring using Deep Learning

The first work published on this topic is the following:

[Denz 20a] F. Denzinger, M. Wels, K. Breininger, M. A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Sühling, and A. Maier. "Automatic CAD-RADS scoring using deep learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 45–54, Springer, 2020

With the main contributions being:

- Introduction of a task-specific data representation for the task at hand.
- Proposal of a bottom-up approach to directly predict the CAD-RADS grade.
- Enhancing the model performance using the stenosis degree and AS score as auxiliary targets.
- Enabling a explainable localization of culprit segments through the network design.
- Providing the first approach to predict all six CAD-RADS grades.

6.2.1 Publication Overview

To aggregate information from the entire coronary artery tree into a patient-wise score, we propose the following workflow. Initially, the coronary centerlines are extracted using the method described in Zheng et al. [Zhen 13] and labeled according to the AHA guidelines using the approach of Gülsün et al. [Guls 14]. In cases where multiple subbranches were assigned the same label, typically occurring for distal vessels, we selected the longest segment. To ensure consistent processing, the segments

were interpolated to a common length. Next, segment-wise features were extracted using the approach presented in our prior work on lesion characterization [Denz 20b].

These segment-wise features were utilized to predict the segment-wise stenosis degree using a multi planar reformation (MPR) approach. It is important to note that, unlike previous work, we utilized binned versions of the stenosis degree based on Cury et al. [Cury 16] instead of the 50 % cutoff. To obtain a patient-wise feature representation, the segment-wise features were combined through a global max pooling operation. This constituent is well motivated by the definition of the CAD-RADS score, which is typically influenced by the most severe lesion score propagated to the patient-level. Moreover, it enables us to identify the segment that contributed the highest activations, providing a localization of the most severely affected segment. This enhances the interpretability of our method. However, due to the weak nature of the segment annotations, we were unable to perform a systematic evaluation of this assumed behavior and only confirmed our intuition qualitatively on a small subset of samples.

Analyzing the results, we observed significant improvements by introducing auxiliary targets for both clinically relevant tasks: rule-out and hold-out. For the rule-out case, we achieved an area under the receiver operating characteristic curve (AUC) of 0.914, while for the hold-out case, the AUC reached 0.923. Additionally, we obtained a Matthew's correlation coefficient (MCC) of 0.424 for the six-class problem.

Automatic CAD-RADS Scoring using Deep Learning

Felix Denzinger[1,2], Michael Wels[2], Katharina Breininger[1], Mehmet A. Gülsün[2], Max Schöbinger[2], Florian André[3], Sebastian Buß[3], Johannes Görich[3], Michael Sühling[2], and Andreas Maier[1]

[1] Pattern Recognition Lab, Universität Erlangen-Nürnberg, Erlangen, Germany
[2] Siemens Healthcare GmbH, Computed Tomography, Forchheim, Germany
[3] Das Radiologische Zentrum - Radiology Center, Sinsheim-Eberbach-Erbach-Walldorf-Heidelberg, Germany

Abstract. Coronary CT angiography (CCTA) has established its role as a non-invasive modality for the diagnosis of coronary artery disease (CAD). The CAD-Reporting and Data System (CAD-RADS) has been developed to standardize communication and aid in decision making based on CCTA findings. The CAD-RADS score is determined by manual assessment of all coronary vessels and the grading of lesions within the coronary artery tree.
We propose a bottom-up approach for fully-automated prediction of this score using deep-learning operating on a segment-wise representation of the coronary arteries. The method relies solely on a prior fully-automated centerline extraction and segment labeling and predicts the segment-wise stenosis degree and the overall calcification grade as auxiliary tasks in a multi-task learning setup.
We evaluate our approach on a data collection consisting of 2,867 patients. On the task of identifying patients with a CAD-RADS score indicating the need for further invasive investigation our approach reaches an area under curve (AUC) of 0.923 and an AUC of 0.914 for determining whether the patient suffers from CAD. This level of performance enables our approach to be used in a fully-automated screening setup or to assist diagnostic CCTA reading, especially due to its neural architecture design – which allows comprehensive predictions.

Keywords: Coronary Artery Disease · Coronary CT Angiography · Deep Learning · Data Representation · CAD-RADS

1 Introduction

Coronary Artery Disease (CAD), which may lead to major adverse events like cardiac infarction or significantly decrease quality of life in the form of coronary ischemia, remains the most common cause of death [7]. Most kinds of CAD result from atherosclerotic plaque deposits aggregating in the vessel wall creating a stenosis, hence narrowing the vessel and obstructing the blood flow. The plaque lesions are categorized by the degree of stenosis into no (0%), minimal (1-24%),

mild (25-49%), moderate (50-69%), severe stenosis (70-99%), and occluded vessel (100%) [2].

Coronary CT Angiography (CCTA) is a common non-invasive rule-out modality for CAD due to its high negative predictive value. In order to standardize communication and guide patient management, the CAD-RADS score based on above mentioned stenosis grades was introduced [2]. It ranges between 0 and 5 and is strongly influenced by the degree of the severest stenosis within a patient. Additionally, this score is influenced by the location of the lesion and includes qualitative assessments based on the experience of the physician, especially in edge-cases.

From a high-level perspective for the case of stable CAD, the resulting patient management decision can be divided into three options: the patient has no CAD and does not need any treatment in the direction of CAD (0), the patient has a non-obstructive CAD (1-2) without need for further investigation, or the patient has an obstructive CAD and should undergo a further functional investigation or direct intervention (3-5).

Therefore, at least these clinical questions need to be answered by an assisting image analysis tool: in the rule-out case, the CAD-RADS 0 score needs to be differentiated from 1-5, and in the hold-out case, the CAD-RADS scores 0-2 need to be differentiated from 3-5. However, prediction on an even finer scale is necessary when the exact required action needs to be identified.

In clinical practice, the assessment of the CAD-RADS score is cumbersome, since the whole coronary tree needs to be assessed and the severest lesion is graded manually based on experience and eyeballing, which is prone to error. Therefore, approaches to ease the workflow and help to detect and grade stenotic lesions have been developed in recent years. Previous approaches focus on detection and quantification of stenoses and are based on the segmentation of the entire coronary tree [5, 10], which is time consuming and often needs manual correction [12].

Recently, deep-learning approaches [6] without the need for a prior segmentation were introduced [1, 11, 14]. These methods operate on multi-planar reformatted (MPR) image stacks which are extracted by interpolating orthogonal planes for each centerline point of the vessel. Approaches for this task include a recurrent convolutional neural network (RCNN) [14], a 2D texture-based multi-view [11], and a 3D CNN approach [1]. A 2D CNN approach, which classifies the whole CCTA volume scaled down and placed in a 2D grid, is described in [8], but might have optimistic results since the training and test splits are described not to be patient-wise.

However, most of the above approaches have the disadvantage of determining the patient score based on single lesions, again introducing a large amount of potential error sources, with no global context incorporated into the decision.

To overcome these pitfalls, we propose a bottom-up approach to directly predict the patient-level CAD-RADS score using a deep-learning based approach that leverages a task-specific hierarchical data representation building up on the coronary tree segments as defined by the American Health Association (AHA)

norm. By having the segment-wise stenosis degree as an additional output and by utilizing a global max pooling operation, which identifies the most relevant features across the whole coronary tree, the network is designed to be comprehensive. Additionally, since all steps in the workflow of our approach can be automated, it can be used for patient screening as well as a preprocessing utility to ease and speed up the clinical workflow.

2 Data

We train and evaluate our methods on a data set consisting of CCTA scans from 2,867 patients collected at a single site.
For each patient, labels regarding the stenosis degree were given on a segment-level as no-stenosis, minimal, mild, moderate, severe or occluded with frequencies of 3,625, 34,889, 4,565, 2,324, 722 and 70 and on patient-level with frequencies of 53, 940, 861, 611, 352 and 50. Furthermore, the CAD-RADS score was annotated on the patient-level with categories 0-5 with frequencies of 436, 584, 873, 568, 348 and 58 [2]. The difference between the patient-wise stenosis degree and the CAD-RADS score can be explained by edge-cases and is especially severe in the CAD-RADS 0 case, since lesions with very minor wall irregularities were classified as minimal according to literature [2]. Additionally, for a subset of 2,828 patients, the Agatston scores were annotated based on additional calcium scoring scans, which were utilized in a binned version according to Rumberger et al. [9] as no, minimal, mild, moderate and severe calcifications with frequencies of 911, 317, 649, 491 and 460.
The data collection did not include patients with stents or bypass grafts. It was split into two parts with two thirds (1,899) used for training and one third (968) used for testing.

3 Methods

Preprocessing For each patient, centerlines are automatically extracted using the algorithm described by Zheng et al. [13] and assigned to the AHA segments [4]. The extracted AHA segment centerlines are used to create MPR image stacks, which are then resized to the mode segment length resulting in a subvolume of size $128 \times 32 \times 32$ for each segment according to Denzinger et al. [3]. Subsequently, the Hounsfield Unit (HU) value range is clipped between -324 and 1,176 HU and normalized to a value range of $[0, 1]$. In order to focus on the more important sections and prevent error propagation from mislabeled AHA-segments, we only select a subset of AHA-segments (RCA_p, RCA_m, RCA_d, LM, LAD_p, LAD_m, LAD_d, LAD_D1, CX_m, CX_d, RAMUS), which were more robustly labeled according to Gülsün et al. [4]. We confirmed to reach similar performance with this subset compared to utilizing all segments in preliminary experiments.

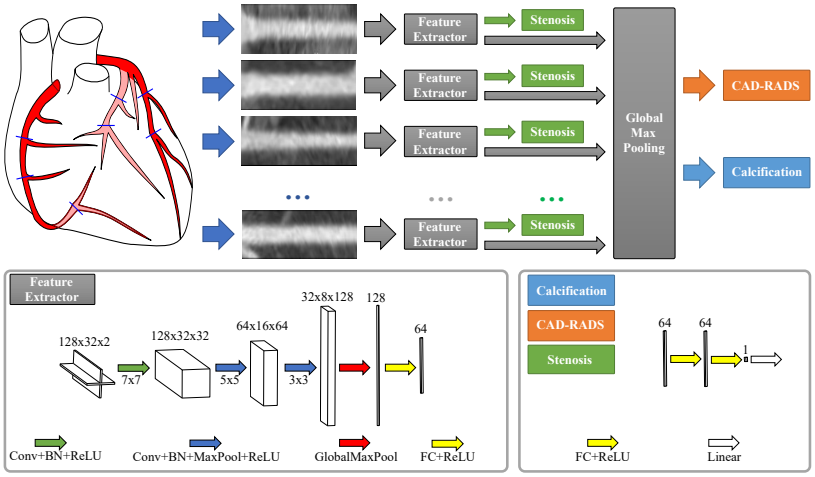

Fig. 1. Model overview (Conv = 2D Convolutional Layer; BN = Batch Normalization Layer; MaxPool = 2D MaxPooling Layer; ReLU = Rectified Linear Unit; FC = Fully-Connected Layer).

Neural Architecture Design The general workflow of the proposed method is outlined in Figure 1. Since our classes are ordered, we reformulate our classification task as a regression problem. This carries the benefit that misclassifications are penalized stronger depending on the class distance, which is convenient since misclassifications between neighboring classes are not as severe for our task. In order to reformat the whole coronary tree in a reasonable representation for neural network training, we divided the whole coronary tree into its sub-segments and extracted straightened MPR volumes. Since we assume all segments to be able to contribute equally, we utilize a feature extractor block with shared weights across all segments to extract spatial features. The feature extractor blocks work on a 2.5D representation utilizing a simple convolutional neural network (CNN) [3]. This architecture choice is motivated by the fact that we strived for simple building blocks to reduce the overall computational effort. Furthermore, we validated in prior experiments that adding additional views or having a feature extractor block similar to the method of [14] did not improve the performance. Since we do not want our model to depend on the location of the stenosis within one segment we choose to decouple the spatial features using a global max pooling operation. A fully-connected layer is used as the last layer of the feature extractor block in order to weight and combine the features such that our different targets can influence each other in the multi-task learning setup. The output of the feature extractor block is then either processed by a stenosis regression block with shared weights across segments to predict the stenosis degree of each segment or the maximum feature responses across all segments are extracted by a global max pooling layer. These global maximum

feature responses are then fed into two further regression blocks for the CAD-RADS and calcification score prediction. This architecture choice is motivated by the definition of the CAD-RADS score as being heavily influenced by the severest lesion. Furthermore, the use of global max pooling allows the network to be more comprehensible since the regions with the highest activations as determined by the network can be displayed to the physician.

Evaluation In order to evaluate the effectiveness of the use of multi-task learning, we evaluated our approach on three different configurations: directly regressing the CAD-RADS score (CAD-RADS), additionally regressing the segment-wise stenosis scores (CAD-RADS + $\asymp$) and also regressing the calcification score (CAD-RADS + $\asymp$ + Ca). Furthermore, to verify whether the global context introduced by our architecture improves the performance, we also evaluate the combination of the feature extractor block and the stenosis regression block with the severest prediction being propagated to the patient-level (Patient-level $\asymp$), which is as close as we can get to related work algorithms with our given labels.
The training set is split into five folds of actual training and validation data (80 %/20 %). The model with the overall lowest loss on the validation set is used as a checkpoint for later evaluation. We choose the Adam optimizer with a learning rate of 0.0001, a batch size of 32 and mean squared error loss for all targets. Furthermore, we utilize data augmentation in the form of rotations around the centerline and minor shifts in x and y direction. In all experiments involving the segment-wise stenosis grade, the feature extractor block is pretrained on the stenosis grade on segment-level before getting integrated into the full model. This is done to condition the feature extractor block towards learning relevant features for the prediction of the stenosis degree. In order to convert our regressed predictions back into classes, we enforce the binned predictions to have the same class distribution as the ground truth labels. The thresholds used for this are calculated on the training set and propagated to the test set.

4 Results

As mentioned in Section 1, most reference approaches perform the classification of the severeness on a per-lesion-level with only Zreik et al. [14] performing an evaluation on the patient-level. However, the severest lesion per patient is not equivalent to the CAD-RADS score and differs especially often in the CAD-RADS 0 case (see Section 2), hence complicating a direct comparison.

CAD-RADS Performance Before analyzing the clinical tasks at hand (rule-out/hold-out), we want to analyze the performance of our approach under different configurations for all six classes. Results for our baseline (severest lesion score as patient score) approach and our full model are given in Fig. 2 and

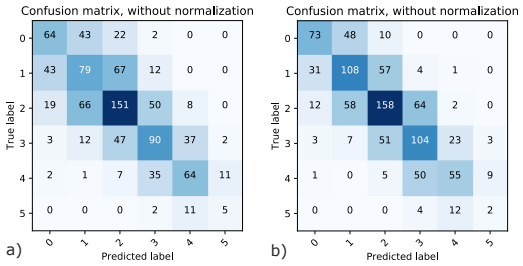

Fig. 2. Example confusion matrices of a single fold on the target of predicting the CAD-RADS using the maximum segment-wise prediction (a) and our proposed multi-task learning procedure (b).

Tab. 1. By leveraging multi-task learning we are able to boost the performance of our approach incrementally (Tab. 1) from an accuracy of 0.810 to 0.840. While the baseline approach performs better compared to direct CAD-RADS scoring without auxiliary targets, we manage to outperform it in the multi-task setup. The biggest performance difference in comparison to the baseline are the lower CAD-RADS scores since in these cases overestimation of single-segment stenoses degrees are especially severe. As displayed in Fig. 2, the hardest class to identify was CAD-RADS 5. An explanation for this is the fact that the centerline extraction fails in the case of occluded vessels. Our method has a low specificity due to the high class imbalance for the single class metrics. Apart from this, most misclassifications are within one class distance, especially in our multi-task learning setup, which is a good feature with respect to the confidence in the network decision.

Table 1. Mean performance on the six class problem of the baseline approach and the three different multi-task learning network configurations ($\asymp$ = segment-wise stenosis grade; Ca = patient-wise calcification grade; MCC = Matthews Correlation Coefficient).

Approach/Metric	Accuracy	Sensitivity	Specificity	MCC
Patient-level $\asymp$	0.825	0.895	0.476	0.371
CAD-RADS	0.810	0.886	0.430	0.316
CAD-RADS + $\asymp$	0.832	0.899	0.496	0.395
CAD-RADS + $\asymp$ + Ca	**0.840**	**0.904**	**0.520**	**0.424**

Rule-out On the task of classifying whether a patient suffers from CAD, we see incremental improvements in the performance of our method with each auxiliary target from an AUC of 0.860 to 0.894 to 0.914 (Fig. 3a and Tab. 2). The performance boost of utilizing the calcification grade can be explained by the fact that

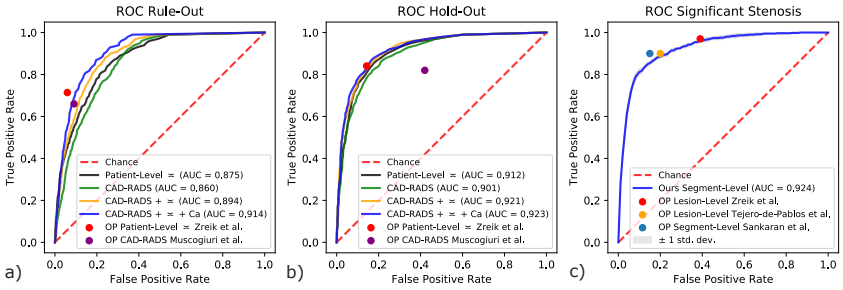

Fig. 3. Results: a) Mean receiver operating characteristic (ROC) curves for the rule-out case. The operating points (OP) of Zreik et al. [14] and Muscogiuri et al. [8] refer to metrics calculated on their data set with Zreik et al. operating on the related task of classifying the severest stenosis degree. b) Mean ROC curves for the hold-out case. c) Mean ROC curve for the classification of significant stenoses. Prediction in Zreik et al. [14] and Tejero-de-Pablos et al. [11] is performed on a per-lesion level and Sankaran et al. [10] utilize the vessel segmentation as additional preprocessing.

patients without CAD should not exhibit any calcifications in the coronary arteries. Also the baseline approach of propagating the severest segment-prediction to the patient-level only reaches an AUC of 0.875 compared to the 0.914 of our full model. Furthermore, there is a severe gap between sensitivity and specificity due to class imbalance. However, as the ROC curve (Fig. 3a) indicates an operating point with both sensitivity and specificity above 0.800 – which is often times required in a clinical setting – can be selected.

Table 2. Results for the rule-out case (predicting CAD-RADS 0 vs 1-5). Results of Zreik et al. [14] refer to the related but different task of predicting the severest stenosis degree on a different data set (abbreviations as in Table 1).

Approach/Metric	Patients	AUC	Accuracy	Sensitivity	Specificity	MCC
Patient-level ≍	955	0.875	0.865	0.508	0.921	0.430
CAD-RADS	955	0.860	0.849	0.489	0.907	0.384
CAD-RADS + ≍	955	0.894	0.875	0.510	0.933	0.456
CAD-RADS + ≍ + Ca	955	**0.914**	**0.888**	**0.532**	**0.945**	**0.504**
Zreik et al. [14]	65	-	0.892	0.714	0.941	0.674
Muscogiuri et al. [8]	284	0.89	0.863	0.660	0.909	0.558

Hold-out In the hold-out case, the use of auxiliary tasks did not boost the performance as much as for the other targets (Fig. 3b and Tab. 3), with the biggest gain caused by adding the segment-wise stenosis degree. However, we outperform our baseline with an AUC, accuracy and MCC of 0.923, 0.860 and 0.692.

Table 3. Results for the hold-out case (predicting CAD-RADS 0-2 vs 3-5). (abbreviations as in Table 1)

Approach/Metric	Patients	AUC	Accuracy	Sensitivity	Specificity	MCC
Patient-level $\asymp$	955	0.912	0.850	0.885	0.781	0.666
CAD-RADS	955	0.901	0.838	0.879	0.759	0.640
CAD-RADS + $\asymp$	955	0.921	0.858	**0.895**	0.787	0.684
CAD-RADS + $\asymp$ + Ca	955	**0.923**	**0.860**	0.891	**0.802**	**0.692**
Zreik et al. [14]	65	-	0.846	0.841	0.857	0.671
Muscogiuri et al. [8]	284	0.78	0.711	0.822	0.583	0.420

Auxiliary Targets For the target of predicting the stenosis degree on a segment-wise level, we reach results comparable to state-of-the-art methods when looking at the binary case of predicting significant stenosis ($>50\,\%$) (Fig. 3c). It should be noted that competing methods are evaluated on different data sets and use labels on lesion-level with defined start and end points, which require a remarkable amount of effort for annotation. Furthermore, our performance on this level enables that segments with the highest score are highlighted in order to aid physicians in their decision making process.
On the task of predicting our calcification grade (as defined in Section 2) we are able to reach a mean accuracy of 0.878.

5 Conclusion

In clinical practice, a standardized way to report CAD from CCTA scans is the CAD-RADS score. To the best of our knowledge – this work presents and evaluates the first approach to directly predict the six class CAD-RADS score using a deep-learning based algorithm. By leveraging two auxiliary tasks – the prediction of the segment-wise stenosis grade and a patient-wise calcification grade – we boosted the performance of our method. The method only relies on a prior centerline extraction and AHA segment label but not on the segmentation of the coronary tree, which is time-consuming to obtain and may need manual correction. Our approach is able to robustly identify patients suffering from CAD (AUC 0.914) or requiring further clinical investigation (AUC 0.923). Segments with severe lesions can be identified by our approach due to the neural architecture design and since we predict segment-wise stenosis with the same network. We validated our approach on a data set of 2,867 patients, a data set considerably larger compared to what has been reported in related work.
Still, the used 2.5D data representation of the single segments may omit some 3D information. We expect this to be successfully addressed by using test augmentation or utilizing additional views in future work. Within this study, it was not possible to apply algorithms defined in related work to our data set, since our stenosis degree labels were segment-wise and not on a per lesion-level. Still, with our experimental design we address this issue in order to allow for a fair comparison. Furthermore, the definition of CAD-RADS also includes report modifiers

related to high-risk plaques, stents and bypass grafts which will be addressed in future work.

Disclaimer The methods and information here are based on research and are not commercially available.

References

1. Candemir, S., et al.: Coronary Artery Classification and Weakly Supervised Abnormality Localization on Coronary CT Angiography with 3-Dimensional Convolutional Neural Networks. arXiv preprint arXiv:1911.13219 (2019)
2. Cury, R.C., et al.: Coronary artery disease-reporting and data system (CAD-RADS): an expert consensus document of SCCT, ACR and NASCI: endorsed by the ACC. JACC: CI **9**(9), 1099–1113 (2016)
3. Denzinger, F., et al.: Deep Learning Algorithms for Coronary Artery Plaque Characterisation from CCTA Scans. In: BVM, LNCS 2020, pp. 193–198. Springer (2020)
4. Gülsün, M.A., Funka-Lea, G., Zheng, Y., Eckert, M.: CTA coronary labeling through efficient geodesics between trees using anatomy priors. In: International Conference on MICCAI. pp. 521–528. Springer (2014)
5. Kirişli, H., et al.: Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography. Medical Image Analysis **17**(8), 859–876 (2013)
6. Maier, A., Syben, C., Lasser, T., Riess, C.: A gentle introduction to deep learning in medical image processing. Zeitschrift für Medizinische Physik **29**(2), 86–101 (2019)
7. Mendis, S., Davis, S., Norrving, B.: Organizational update: the World Health Organization global status report on noncommunicable diseases 2014. Stroke **46**(5), e121–e122 (2015)
8. Muscogiuri, G., et al.: Performance of a deep learning algorithm for the evaluation of CAD-RADS classification with CCTA. Atherosclerosis **294**, 25–32 (2020)
9. Rumberger, J., Kaufman, L.: A rosetta stone for coronary calcium risk stratification: agatston, volume, and mass scores in 11,490 individuals. American Journal of Roentgenology **181**(3), 743–748 (2003)
10. Sankaran, S., et al.: Hale: Healthy area of lumen estimation for vessel stenosis quantification. In: International Conference on MICCAI. pp. 380–387. Springer (2016)
11. Tejero-de-Pablos, A., et al.: Texture-Based Classification of Significant Stenosis in CCTA Multi-view Images of Coronary Arteries. In: MICCAI. pp. 732–740. Springer (2019)
12. Wels, M., Lades, F., Hopfgartner, C., Schwemmer, C., Sühling, M.: Intuitive and accurate patient-specific coronary tree modeling from cardiac computed-tomography angiography. In: The 3rd interactive MIC Workshop. pp. 86–93 (2016)
13. Zheng, Y., Tek, H., Funka-Lea, G.: Robust and accurate coronary artery centerline extraction in CTA by combining model-driven and data-driven approaches. In: MICCAI. pp. 74–81. Springer (2013)
14. Zreik, M., et al.: A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography. IEEE Transactions on Medical Imaging **38**(7), 1588–1598 (2018)

6.3 CAD-RADS Scoring using Deep Learning and Task-Specific Centerline Labeling

Building up on our previously presented method, we published:

[Denz 21b] F. Denzinger, M. Wels, O. Taubmann, M. A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Suehling, and A. Maier. "CAD-RADS Scoring using Deep Learning and Task-Specific Centerline Labeling". In: *Medical Imaging with Deep Learning*, 2021

Here, our contibutions include:

- Proposal of a task-specific heuristic centerline labeling.
- Evaluation of a severity-based label encoding leveraging the ordinal nature of the classification task.
- Use of test time augmentation (TTA) and model ensembling to enhance the robustness of the final approach.

6.3.1 Publication Overview

Upon examining our initial method, we identify certain weaknesses primarily associated with the pre-processing pipeline. While we have confidence in the centerline extraction algorithm, we notice suboptimal performance in a small number of cases for the centerline labeling. In these instances, some subbranches are missing or the distal subbranches are labeled at more proximal points than usual. While this can occasionally be accurate, it results in significant variability in the lengths of individual segments. To address this issue and simplify the processing, we propose a straightforward heuristic approach for centerline labeling. This approach involves dividing the centerlines into the three main branches using a simple rule set and selecting the longest branches for each main branch as the primary branches. These primary branches are then divided into segments of equal length. By employing this approach, we provide the downstream artificial neural network (ANN) with a set of segments that do not suffer from interpolation inhomogeneities. Additionally, the segments are now comparable across patients, ensuring that the same anatomical regions are assessed.

Another area for improvement we identified was the encoding of labels for network training. Previously, the problem was formulated as a regression task with subsequent threshold optimization to obtain individual class predictions. While this enforced

predictions to follow the ordering within the data collection, it did not account for the actual classes. To enhance this aspect, we propose the utilization of a rank-based encoding.

Furthermore, we incorporated additional concepts in this study, namely TTA and model ensembling, as they are known to enhance the robustness of DL-based approaches. TTA had already proved useful in our prior work, which served as the basis for our model architecture [Denz 20b]. In the current publication, we took a step further and eliminated the orthogonal view, reducing the network's footprint, as it appeared that the task at hand did not benefit from it.

The final results demonstrated that all the modifications yielded improvements across all tasks. For the rule-out and hold-out cases, we achieved an AUC of 0.942 and 0.950, respectively. Additionally, for the six-class problem, an MCC of 0.493 was obtained. These outcomes highlight the efficacy of the proposed changes and their positive impact on the overall performance of the system.

Proceedings of Machine Learning Research 172:1–10, 2022 Full Paper – MIDL 2022

CAD-RADS Scoring using Deep Learning and Task-Specific Centerline Labeling

Felix Denzinger[1,2] FELIX.DENZINGER@FAU.DE
[1] *Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[2] *Siemens Healthcare GmbH, Computed Tomography, Forchheim, Germany*

Michael Wels[2]
Oliver Taubmann[2]
Mehmet A. Gülsün[2]
Max Schöbinger[2]
Florian André[3]
[3] *Das Radiologische Zentrum - Radiology Center, Sinsheim-Eberbach-Erbach-Walldorf-Heidelberg, Germany*

Sebastian J. Buss[3]
Johannes Görich[3]
Michael Sühling[2]
Andreas Maier[1]
Katharina Breininger[4]
[4] *Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*

Abstract

With coronary artery disease (CAD) persisting to be one of the leading causes of death worldwide, interest in supporting physicians with algorithms to speed up and improve diagnosis is high. In clinical practice, the severeness of CAD is often assessed with a coronary CT angiography (CCTA) scan and manually graded with the CAD-Reporting and Data System (CAD-RADS) score. The clinical questions this score assesses are whether patients have CAD or not (rule-out) and whether they have severe CAD or not (hold-out). In this work, we reach new state-of-the-art performance for automatic CAD-RADS scoring. We propose using severity-based label encoding, test time augmentation (TTA) and model ensembling for a task-specific deep learning architecture. Furthermore, we introduce a novel task- and model-specific, heuristic coronary segment labeling, which subdivides coronary trees into consistent parts across patients. It is fast, robust, and easy to implement. We were able to raise the previously reported area under the receiver operating characteristic curve (AUC) from 0.914 to **0.942** in the rule-out and from 0.921 to **0.950** in the hold-out task respectively.

Keywords: Coronary Artery Disease, Coronary CT Angiography, Deep Learning, Ensembling, CAD-RADS, Coronary Artery Labeling

1. Introduction

Worldwide, coronary artery disease (CAD) still is the leading cause of death (Roth et al., 2020), thus impacting the lives of many. Therefore, developing algorithms to support physicians with the diagnosis is of high interest. These algorithms may serve as a second reader

to ensure that no aspect is missed or to point the physician to areas of interest, thus speeding up the workflow.
CAD is predominantly linked to atherosclerotic plaque deposits aggregating within the vessel wall (Fuster et al., 1992). The degree of vessel narrowing – also called stenosis – caused by such a plaque deposit is an essential piece of information regarding patient risk and can be obtained using a coronary CT angiography (CCTA) scan. To report findings, assess patients' general condition, and to guide the clinical workflow the coronary artery disease-reporting and diagnosis system (CAD-RADS) score was introduced (Cury et al., 2016). This score is usually determined through a manual assessment by a human reader scoring the whole coronary vessel tree. It consists of six grades ranging from 0 to 5, where 0 refers to "no CAD present", 1-2 to "non-obstructive CAD present" and 3-5 to "obstructive CAD present", with a rising severeness within this grouping. Hence, primary clinical questions of interest are whether patients do have CAD or not (rule-out) and whether they suffer from obstructive CAD and therefore should undergo further (invasive) assessment including potential immediate revascularization or not (hold-out). However, this manual grading is time-consuming and reader/experience dependent (Razek et al., 2018; Maroules et al., 2018; Hu et al., 2021). Therefore, introducing decision support algorithms for this task is of high interest.As related work regarding this task is sparse, we discuss work on the related task of predicting severe stenosis degree. Algorithms performing this task can be divided into lesion-wise, and branch-wise.Lesion-wise algorithms focus mainly on the task of detecting and (separately) scoring one or multiple plaque deposits within the whole coronary vessel tree. Most of these approaches work on multi-planar reformatted (MPR) volumes created by interpolating orthogonal planes for each vessel centerline point. Commonly, these approaches are based on recurrent convolutional neural networks (RCNN) (Zreik et al., 2018; Denzinger et al., 2019; Ma et al., 2021). For these, a series of overlapping cubes along the centerline dimension is used, from which spatial features are extracted using a 3D convolutional neural network (CNN) at each position. The resulting feature sequence is analyzed using a recurrent neural network (RNN) (Zreik et al., 2018) or combined using a transformer module (Ma et al., 2021). A branch-wise approach presented by (Candemir et al., 2020) utilizes a 3D CNN which takes whole coronary branches in MPR format as inputs. Disadvantages of both lesion- and branch-wise approaches are that errors on lesion-/branch-level are directly propagated to patient-level and that only local information is included in the network prediction. A case-wise CAD severity score is the Agatston score (Agatston et al., 1990), which in principle assesses the overall calcified plaque burden of a patient from non-contrast CT scans. This score can also be determined using machine learning methods (Wolterink et al., 2014; Lessmann et al., 2017; Cano-Espinosa et al., 2018). Our group recently proposed a case-wise approach to determine the CAD-RADS score (Denzinger et al., 2020b). It uses a hierarchical data representation of the whole coronary tree based on its anatomical sub-segments. For each of these sub-segments, features are extracted from the MPR volume stack with a CNN and combined with a global max pooling layer to predict the case-wise score. Based on the architecture and concepts presented in our previous work (Denzinger et al., 2020b), we present a more robust, streamlined and reproducible pipeline. Specifically, to ease reproducibility and simplify the pre-processing pipeline of our work we propose an architecture- and task-specific heuristic centerline labeling. Moreover,

we are leveraging the use of a severity-based label encoding, test time augmentation (TTA), model ensembling and reduced input dimensionality.

2. Data

Data is provided from a single site with CCTA scans acquired with the same scanner type. The number of patients (and samples) included is 2,902 with a fixed split of 1,926 used for training and 976 for testing. Within the test set, 131 patients have no CAD, 499 patients have non-obstructive CAD and 346 patients have obstructive CAD. The pre-processing is conducted as follows: after extracting the coronary centerlines using the method of (Zheng et al., 2013), MPR image stacks are extracted by interpolating planes orthogonal to the centerlines with a spacing of $(0.33 \times 0.33)\,\mathrm{mm}^2$ and a field of view (FOV) of $12 \times 12\,\mathrm{mm}^2$ for each centerline point with centerline points placed 0.25 mm apart. For these MPR image stacks, the Hounsfield unit (HU) value range is clipped to lie between -300 HU and 1,024 HU with the resulting values being rescaled to a value range between 0 and 1.

3. Methods

3.1. Architecture

An overview of the used deep learning architecture is presented in Fig. 1, including an explanation of the individual steps.

3.2. Proposed Extensions

As the input for this network is either one or two orthogonal longitudinal views cut from the MPR volume stack at a specific angle α for each subsegment (cf. Fig. 1), the information used to predict the CAD-RADS score may vary. Therefore, the prediction may not be consistent with different angles, which it should be, given that for all angles the same biological information should be assessed. Our group showed in previous work (Denzinger et al., 2020a) that this problem can be partly solved by adding a second orthogonal view which still leaves some leeway for suboptimal angles especially when only one angle is considered during inference. To overcome this we leverage TTA averaging predictions for 16 views extracted for equally distributed angles between $[0, \pi]$ with the same angle for all segments. As the whole vessel information should be covered with this strategy, we additionally evaluate whether a single longitudinal view instead of two orthogonal longitudinal views suffices. Also, we propose to use model ensembling to lower uncertainty introduced by the network training converging to different local optima. In our prior work (Denzinger et al., 2020b), the prediction of the CAD-RADS score is transformed from a classification to a regression task and the network trained with a mean squared error (MSE) loss. This leads to all classes being weighted equally and the loss not depending on the individual class and how well this class has been learned already. To address this we suggest to use the following label encoding (Niu et al., 2016): $y_i^k = 1$ if $i \leq k$, $y_i^k = 0$ otherwise. Therefore, label vectors $\mathbf{y}^k$ belonging to class k are created, with i denoting the index of the entry in the label vector (e.g. CAD-RADS 2 is encoded as (1,1,1,0,0,0)). With this, we transform the regression task to a multi-label problem, which enables the use of a cross-entropy loss with

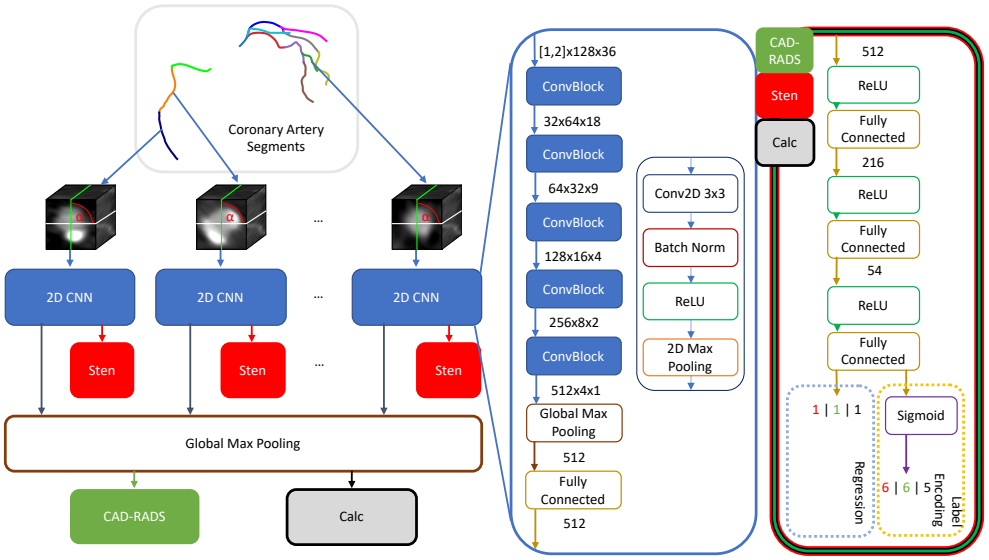

Figure 1: Overview of the architecture. For each labeled subsegment an MPR volume stack is computed and for one arbitrary angle α around the centerline, a longitudinal slice or two orthogonal longitudinal slices are extracted. The slices of all segments are fed into the same 2D CNN. The resulting feature representation is further processed by a multi-layer perceptron (MLP) for each segment to classify the stenosis grade and globally max pooled. The global feature representation is fed into two MLPs predicting the overall calcification (denoted as Calc and determined as a binned version of the Agatston score according to (Rumberger and Kaufman, 2003)) and the CAD-RADS grade. The output of the network is either one scalar value in case of regression or 5-6 sigmoidal outputs in case the labels are encoded as described in Section. 3.2.

sigmoidal predictions. During inference the raw predictions are summed over all outputs to get a cumulative probability and binned according to (Denzinger et al., 2020b).

3.3. Centerline Labeling

Furthermore, in the pipeline described in Reference (Denzinger et al., 2020b), the coronary tree was subdivided using the method proposed by (Gülsün et al., 2014) and the resulting segments were interpolated to one common length. With this a reasonable input to the network is obtained which may, however, yield obscured segments. Moreover, the extracted coronary tree usually exhibits more centerlines than defined in literature, since also small side branches are found by the centerline extraction algorithm of (Zheng et al., 2013). Furthermore, distal parts are usually less important and if a stenosis is present there it has less influence on thrombus formation or myocardial ischemia. Therefore, these should not necessarily have an impact on network prediction. Furthermore, even if the segment labels determined with the method of (Gülsün et al., 2014) are anatomically correct – which is

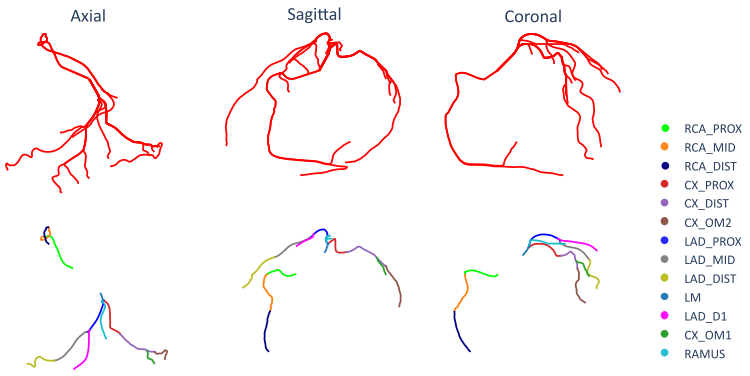

Figure 2: Centerlines before (top) and after (bottom) labeling. Note that centerline points inside the aorta originate from our data format and do not need to be labeled. Detected centerlines include the proximal, mid and distal part of the right coronary artery RCA (RCA_{prox}, RCA_{mid} and RCA_{dist}), the left main segment (LM), the proximal, mid and distal part of the left artery descending (LAD) and left circumflex artery (CX) named LAD_{prox}, LAD_{mid}, LAD_{dist} and CX_{prox}, CX_{dist}, CX_{OM2}, respectively, and the obtuse marginal (OM) artery of the CX CX_{OM1} and the diagonal segment of the LAD, LAD_{D1}.

not always guaranteed – the segment image information is not directly transferable between patients due to the different segment lengths and potentially different supplied heart regions. We therefore propose an heuristic centerline labeling approach to solve previously mentioned problems with following notation [1]: let $\mathbf{C}$ be a set of centerlines C consisting of centerline points $\mathbf{c} \in \mathbb{R}^3$. $\mathbf{c}_0$ is the first point of each centerline, which in our centerline format is always the center of the aorta with the first centerline points connecting the center of the aorta with the respective ostia. All centerlines end at their respective most distal point $\mathbf{c}_{n_c}$. This format leads to high redundancy in the centerlines with proximal parts often overlapping. An example of this and the abbreviations for the different segments are included in Fig. 2. Our heuristic pipeline is defined as follows: the set of centerlines can be subdivided into left $\mathbf{C}_l$ and right $\mathbf{C}_r$ centerlines by looking at their world coordinate direction starting from the center of the aorta. If hypothetically a different centerline-extraction algorithm outputs centerlines starting from the ostia, this initial step could be skipped. For the right centerline tree the longest segment C_r^* is selected and, starting from the ostium, three subsequent segments of length 32 mm each are labelled as RCA_{prox}, RCA_{mid} and RCA_{dist}, while the remaining vessel is excluded. For the left coronary tree, the bifurcation point $\mathbf{c}_b$ between the LAD and CX needs to be determined first. We detect $\mathbf{c}_b$ as the point where the centerlines of the left tree split most frequently. The LM is consequently labeled as the segment between the left ostium and $\mathbf{c}_b$. From $\mathbf{c}_b$ we calculate the directions of all centerlines containing this point as $\mathbf{c}_{b+10} - \mathbf{c}_b$. If there are **two** unique directions, the rightmost centerlines with this direction are defined as the LAD branch and the leftmost as

1. Code available at https://github.com/fdenz/HeuristicCLLabeling

the CX branch. If there are **three** unique directions the branch between the others is labeled as RAMUS intermedius, which does not exist for all patients. The longest centerlines C^*_{LAD} and C^*_{CX} of the LAD and CX are divided into LAD_{prox}, LAD_{mid}, LAD_{dist} and CX_{prox}, CX_{dist}, CX_{OM2} respectively to obtain segments of lengths 32 mm. Furthermore, for LAD and CX the centerlines C'_{LAD} and C'_{CX} which have the longest non-overlapping part to C^*_{LAD} and C^*_{CX} are selected. The 32 mm segments starting from the bifurcation between C'_{LAD} / C'_{CX} and C^*_{LAD} / C^*_{CX} respectively are labeled as LAD_{D1} and CX_{OM1}. As the described heuristic approach does not aim to be absolutely anatomically correct and relies only on a small set of rules, it is consistent by design. Furthermore, it extracts segments of the same length, which eases the network training when it compares the segments of different patients. On the other hand, bifurcations do not only occur on the start and end of the segments, but also in the middle of the segment which leads to more diverse training data. Also, and maybe most importantly, it is simple and fast (around 350ms with a Intel(R) Xeon(R) CPU E5-2640 CPU).

3.4. Evaluation

For the evaluation, we keep our test set fixed while splitting our training data into five parts of approximately equal size in a stratified manner. We then use four of these parts as training and one as a validation set for five folds with the best model for each training with respect to the validation CAD-RADS score loss saved for evaluation. This setting is repeated five times for different seeds and splits for a total of 25 trained models for all configurations. Further hyperparameters were a stochastic gradient descent optimizer with a learning rate of 0.007/0.0007 for the label encoding/regression task respectively and a momentum of 0.99. We evaluate our different additions in form of an incremental study. First, the centerline labeling of the original approach is replaced with the one described in the section above. As the prediction of the network depends on the angle selected we will include the average results over all angles, the initial angle, and the angle with the retrospectively highest performance. Next, we use TTA taking the mean prediction over 16 angles equally distributed between $[0, \pi]$ with the same angle applied to all segments. This is followed by ensembling models and taking either the average prediction over the five folds of one seed or all 25 models. Finally, the label encoding is added, before testing whether a single view suffices.

4. Results and Discussion

As we have an ordinal classification task and are able to adapt the threshold depending on the desired ratio of sensitivity and specificity, we consider the area under the receiver operating curve (AUC) to be the most important metric. In general, we can see an incremental increase in performance with each improvement for the clinical question of rule-out (Table 1), hold-out (Table 2) and for predicting all six CAD-RADS grades (Table 3).
As we previously only reported the metrics for the views at a single angle in Reference (Denzinger et al., 2020b), it is hard to select which angle to choose for comparison. This task is also impacted by the fact that results differ at different selected angles. When averaging over all evaluated angles we get a mean AUC of 0.913 compared to 0.914 for the rule-out and 0.933 compared to a baseline of 0.923 for the hold-out case. However, looking at the

Config/Metric	AUC	ACC	Sens	Spec	MCC
Baseline	0.914	0.888	0.532	0.945	0.504
+ ~~TTA~~ E_1 ~~LE~~ 0	0.917±0.008	0.884±0.006	0.569±0.023	0.933±0.003	0.503±0.027
+ ~~TTA~~ E_1 ~~LE~~ *	0.917±0.008	0.886±0.008	0.574±0.031	0.934±0.004	0.508±0.036
+ ~~TTA~~ E_1 ~~LE~~ ∀	0.913±0.006	0.880±0.004	0.555±0.016	0.931±0.002	0.486±0.019
+ TTA E_1 ~~LE~~	0.924±0.005	0.887±0.007	0.578±0.026	0.935±0.004	0.512±0.030
+ TTA E_5 ~~LE~~	0.932±0.001	0.890±0.002	0.591±0.007	0.937±0.001	0.527±0.008
+ TTA E_{25} ~~LE~~	0.934	0.891	0.595	0.937	0.533
+ TTA E_{25} LE	0.941	0.895	0.611	0.940	0.550
− TTA E_{25} LE	**0.942**	**0.912**	**0.672**	**0.949**	**0.621**

Table 1: Performance for the **rule-out task** for the different model configurations. Metrics are: the area under the receiver operating curve (AUC), accuracy (ACC), sensitivity (Sens), specificity (Spec), and Matthews correlation coefficient (MCC). "+/−" denotes whether two orthogonal or one single longitudinal view is fed into the CNN, "~~TTA~~/TTA" whether TTA is used, "E_i" the number of models ensembled, "~~LE~~/LE" whether labels are encoded as described in Section. 3.2 and "0/*/∀" whether the views extracted for the first, retrospectively best or all evaluated angles were considered. Baseline refers to the results reported in Reference (Denzinger et al., 2020b).

Config/Metric	AUC	ACC	Sens	Spec	MCC
Baseline	0.923	0.860	0.891	0.802	0.692
+ ~~TTA~~ E_1 ~~LE~~ 0	0.932±0.003	0.854±0.006	0.887±0.005	0.794±0.009	0.680±0.014
+ ~~TTA~~ E_1 ~~LE~~ *	0.937±0.004	0.860±0.007	0.892±0.005	0.803±0.009	0.695±0.015
+ ~~TTA~~ E_1 ~~LE~~ ∀	0.933±0.004	0.856±0.004	0.888±0.003	0.797±0.006	0.686±0.010
+ TTA E_1 ~~LE~~	0.940±0.004	0.861±0.005	0.893±0.003	0.804±0.007	0.697±0.011
+ TTA E_5 ~~LE~~	0.943±0.000	0.860±0.002	0.892±0.001	0.803±0.003	0.695±0.005
+ TTA E_{25} ~~LE~~	0.943	0.861	0.892	0.803	0.696
+ TTA E_{25} LE	0.944	0.861	0.892	0.803	0.696
− TTA E_{25} LE	**0.950**	**0.877**	**0.905**	**0.827**	**0.731**

Table 2: Performance for the **hold-out question** for the different model configurations. Abbreviations as in Table 1.

angle with the best overall performance or the initial angle as an example, the performance is better than the baseline performance. This also nicely demonstrates why TTA is crucial. With TTA, a clear improvement in general performance is observed. This is easily explained by the fact that lesions cannot be missed by an unfortunate angle anymore. Ensembling multiple models leads to another performance boost, with an obvious improvement in stability when observing the decrease in standard deviation as the metric. Our proposed label encoding results in no improvement for the hold-out case, as the class balance is less severe in this case. However, for the rule-out case, an improvement from an AUC of 0.934 to

Config/Metric	ACC	Sens	Spec	MCC
Baseline	0.840	0.904	0.520	0.424
+ ~~TTA~~ E_1 ~~LE~~ 0/*	0.839±0.005	0.904±0.003	0.518±0.017	0.422±0.021
+ ~~TTA~~ E_1 ~~LE~~ ∀	0.834±0.004	0.900±0.002	0.504±0.014	0.405±0.017
+ TTA E_1 ~~LE~~	0.841±0.005	0.904±0.003	0.522±0.017	0.426±0.020
+ TTA E_5 ~~LE~~	0.842±0.001	0.905±0.000	0.525±0.004	0.430±0.005
+ TTA E_{25} ~~LE~~	0.844	0.906	0.532	0.438
+ TTA E_{25} LE	0.845	0.907	0.535	0.442
− TTA E_{25} LE	**0.859**	**0.916**	**0.578**	**0.493**

Table 3: Performance for the **six-class problem** averaged over all classes for the different model configurations. Abbreviations as in Table 1.

0.941 is observed. This illustrates that this change improves differentiation of less frequent classes. Finally, we tested decreasing the dimensionality by only having a single longitudinal view combined with TTA for each segment as an input for the network. This yielded far better results for all targets. A possible explanation for this may be the increased training stability that we observed and that the same information is fed to the system due to TTA. Moreover, beforehand the ordering of the two orthogonal longitudinal slices led to different results as different features were extracted for each, which should not be of relevance for the targets at hand. Especially the metrics for the six-class problem benefited the most from this change.

5. Conclusion

In this paper, we improve the automatic deep learning-based assessment of patients regarding the CAD-RADS score. We propose the use of TTA, model ensembling, task-specific label encoding, and reduced model input dimensionality for this task. Moreover, we introduce a novel task-specific heuristic centerline labeling approach, which by itself does neither lead to improved nor worse performance. However, it is easy to implement and makes the whole model pipeline easier to reproduce, while being theoretically more robust to technical variations due to its heuristic nature. Overall, we improve previously reported performance on the data set at hand: the accuracy for the six-class problem is increased to 0.859 from 0.840 and the AUC for the rule-out case to 0.942 from 0.914. For the hold-out case, we were able to reach an AUC of 0.950 compared to a previously reported 0.923. Further steps for this method are to apply it to data at different sites and/or scanner types.

Disclaimer: The methods and information here are based on research and are not commercially available.

Acknowledgement: K.B. gratefully acknowledges the support of the project "Dhip campus - bavarian aim".

References

Arthur S Agatston, Warren R Janowitz, Frank J Hildner, Noel R Zusmer, Manuel Viamonte, and Robert Detrano. Quantification of coronary artery calcium using ultrafast computed tomography. Journal of the American College of Cardiology, 15(4):827–832, 1990.

Sema Candemir et al. Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network. Computerized Medical Imaging and Graphics, 83:101721, 2020.

Carlos Cano-Espinosa, Germán González, George R Washko, Miguel Cazorla, and Raúl San José Estépar. Automated Agatston score computation in non-ECG gated CT scans using deep learning. In Medical Imaging 2018: Image Processing, volume 10574, page 105742K. International Society for Optics and Photonics, 2018.

Ricardo C Cury et al. Coronary artery disease-reporting and data system (CAD-RADS): an expert consensus document of SCCT, ACR and NASCI: endorsed by the ACC. JACC: CI, 9(9):1099–1113, 2016.

Felix Denzinger et al. Coronary Artery Plaque Characterization from CCTA Scans Using Deep Learning and Radiomics. In International Conference on MICCAI, pages 593–601. Springer, 2019.

Felix Denzinger et al. Deep Learning Algorithms for Coronary Artery Plaque Characterisation from CCTA Scans. In BVM, LNCS 2020, pages 193–198. Springer, 2020a.

Felix Denzinger et al. Automatic CAD-RADS Scoring Using Deep Learning. In International Conference on MICCAI, pages 45–54. Springer, 2020b.

Valentin Fuster, Lina Badimon, Juan J Badimon, and James H Chesebro. The pathogenesis of coronary artery disease and the acute coronary syndromes. New England Journal of Medicine, 326(5):310–318, 1992.

Mehmet A Gülsün, Gareth Funka-Lea, Yefeng Zheng, and Matthias Eckert. CTA coronary labeling through efficient geodesics between trees using anatomy priors. In International Conference on MICCAI, pages 521–528. Springer, 2014.

Jiun-Yiing Hu et al. Interobserver reliability of the coronary artery disease reporting and data system in clinical practice. Journal of thoracic imaging, 36(2):95–101, 2021.

Nikolas Lessmann, Bram van Ginneken, Majd Zreik, Pim A de Jong, Bob D de Vos, Max A Viergever, and Ivana Išgum. Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions. IEEE transactions on medical imaging, 37(2): 615–625, 2017.

Xinghua Ma, Gongning Luo, Wei Wang, and Kuanquan Wang. Transformer Network for Significant Stenosis Detection in CCTA of Coronary Arteries. In International Conference on MICCAI, pages 516–525. Springer, 2021.

Christopher D Maroules et al. Coronary artery disease reporting and data system (CAD-RADSTM): Inter-observer agreement for assessment categories and modifiers. Journal of cardiovascular computed tomography, 12(2):125–130, 2018.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal Regression with Multiple Output CNN for Age Estimation. In Proceedings of the IEEE conference on CVPR, pages 4920–4928, 2016.

Ahmed Abdel Khalek Abdel Razek et al. Inter-observer agreement of the coronary artery disease reporting and data system (CAD-RADSTM) in patients with stable chest pain. Polish journal of radiology, 83:e151, 2018.

Gregory Roth et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. Journal of the American College of Cardiology, 76 (25):2982–3021, 2020.

John Rumberger and Leon Kaufman. A rosetta stone for coronary calcium risk stratification: agatston, volume, and mass scores in 11,490 individuals. American Journal of Roentgenology, 181(3):743–748, 2003.

Jelmer M Wolterink, Tim Leiner, Richard AP Takx, Max A Viergever, and Ivana Išgum. An automatic machine learning system for coronary calcium scoring in clinical non-contrast enhanced, ECG-triggered cardiac CT. In Medical Imaging 2014: Computer-Aided Diagnosis, volume 9035, pages 110–117. SPIE, 2014.

Yefeng Zheng, Huseyin Tek, and Gareth Funka-Lea. Robust and accurate coronary artery centerline extraction in CTA by combining model-driven and data-driven approaches. In MICCAI, pages 74–81. Springer, 2013.

Majd Zreik et al. A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography. IEEE Transactions on Medical Imaging, 38(7):1588–1598, 2018.

6.4 Discussion

With both publications presented, it is crucial to provide a broader context by examining related work and discussing the advantages and disadvantages of our approaches.

Overall, our method demonstrates high performance and consists of a clinically motivated data representation and architecture design. In comparing with other approaches, we can analyze their similarities and differences. For instance, Paul et al. [Paul 22] also focus on patient-level CAD-RADS grading, specifically the rule-out and hold-out differentiation. They employ curved planar reformation (CPR) views instead of the stretched MPR format used in our work. While their method shares similarities with ours, such as the use of TTA to incorporate differing angles, we introduce a heuristic centerline labeling approach that includes the second largest subbranches for the left artery descending (LAD), circumflex artery (CX), and the ramus intermedius (RI) in our input, while Paul et al. only focus on the three main branches. This inclusion allows for potential detection of lesions in these branches, which could be missed by Paul et al.'s approach. However, the impact of such lesions may be less significant as most relevant lesions tend to be in the proximal regions and/or in the primary branches. Hence, we acknowledged this limitation and maintained a similar drawback in our approach.

Jin et al.[Jin 22] proposed an approach that solely utilizes the coronary CT angiography (CCTA) volume as input for five-class CAD-RADS grading, excluding CAD-RADS 0. In contrast to the centerline extraction we leveraqge, they employ a object detection and segmentation step. While their approach provides a definite lesion localization and explainable predictions, it may also introduce error propagation throughout the pipeline. In our approach, although we lack distinct lesion localization, the most important segment for the final prediction can be determined through the global max pooling, and the angle with the highest score from TTA can be extracted to present the most impactful 2D input to the physician. This enables an understanding of the network's reasoning behind the final prediction.

As another approach, Penso et al.[Pens 23] also consider the main branches as input, leading to similar limitations as Paul et al.[Paul 22]. Their approach primarily focuses on branch-wise classification and propagates the most severe grading to the patient level. This procedure can be applied to other methods discussed in Chapter 4 and Chapter 5. Contrary to this, our approach incorporates a global max pooling operation at the feature level, allowing for information aggregation across multiple segments and a holistic and comparative aspect. However, it is important to note that

this reasoning has not been validated. Nonetheless, we conducted experiments where only the feature representation of the segment with the highest overall activation was propagated to the patient level, and the results showed a worse performance, confirming our initial reasoning.

In summary, the comparison with related work highlights the unique aspects and contributions of our approach. While other studies have explored patient-level CAD-RADS grading using different techniques, such as alternative image views or object detection and segmentation, our method stands out for its clinically motivated data representation and architecture design. We address limitations in centerline labeling, incorporate additional sub-branches for analysis, and provide interpretability through feature pooling and angle extraction. These modifications lead to a comprehensive evaluation of the CAD-RADS grading task.

CHAPTER 7

Clinical Applicability

With a well-performing coronary artery disease (CAD) classification approach presented, the open question regarding generalizability of the method with respect to differing image formation parameters is evaluated in this chapter. Furthermore, efforts were made to introduce and evaluate a strategy for handling data where annotators were not able to label all samples confidently. Here, also learning with abstention was explored with a novel quantile-based abstention strategy.

7.1 Influence of Scan Parameters to Deep Learning-based CAD-RADS Scoring

Taking the approach described in Section 6.3.2 as a basis we released the following research:

[Denz 23a] F. Denzinger, M. Wels, K. Breininger, O. Taubmann, A. Mühlberg, T. Allmendinger, M. A. Gülsün, M. Schöbinger, F. André, S. J. Buss, J. Görich, M. Sühling, and A. Maier. "How scan parameter choice affects deep learning-based coronary artery disease assessment from computed tomography". *Scientific Reports*, Vol. 13, No. 1, p. 2563, 2023

The main novelties of this work include:

- Evaluating deep learning (DL)-based coronary artery disease-reporting data system (CAD-RADS) scoring with respect to the influence of selected reconstruction parameters.
- Being the first approach to perform this analysis on this task with paired data.
- Examining how the pre-processing steps influence the final prediction.

- Drawing conclusions, which computed tomography (CT) reconstruction parameters need to be treated with caution.

7.1.1 Publication Overview

Throughout this thesis, it is emphasized that CAD scoring using coronary CT angiography (CCTA) data involves multiple interconnected steps, from scan acquisition to the final prediction. The image formation process itself plays a crucial role in influencing the subsequent pipeline steps and the overall CAD grading outcome. In an ideal scenario, the prediction of an artificial neural network (ANN) for CAD grading would rely solely on the underlying biological information. To assess the need for disentangling biological and technical information, we conduct a study using the following setting. For a subset of data we utilize in Chapter 6, we obtain the raw acquisition data and reconstruct it with various CT reconstruction parameters often altered in clinical practice. These include the advanced modeled iterative reconstruction (ADMIRE) strength, stacking strategy, and reconstruction kernel, which are detailed in Chapter 2. We then examine the change in prediction resulting from these parameter alterations.

By conducting this analysis with paired data, we gain a more precise understanding of causality compared to statistical observations over a larger data population. We also separate the influence of the image formation parameters from the ANN component by propagating the centerline extraction results of a default configuration to other parameter variations. Our findings revealed an intriguing pattern for most parameter configurations, wherein the overall performance did not exhibit significant changes, but individual patient grades were affected. This raises ethical considerations regarding the importance of robust global performance versus patient-wise accuracy. Notably, we obtained definitive conclusions regarding the choice of the stacking strategy, where the true stack configuration demonstrated significantly lower performance for the rule-out case. Additionally, sharper reconstruction kernels resulted in more individual class changes. Furthermore, we observed that the centerline extraction step in our pipeline was also influenced by changes in image formation parameters, indicating the need for further research in this area.

7.1.2 Discussion

After defining the scope and presenting our work's results, it is essential to place it within a broader context. The initial motivation behind this research was to address the question of whether it is necessary to disentangle biological and technical information. As demonstrated, there is a significant variation in patient-wise CAD-

RADS gradings due to changes in image formation, which is an undesirable behavior. Therefore, it is crucial to develop techniques to overcome this challenge.

One possible approach is to employ a training setting where the feature representations for different parameter configurations are enforced to be similar. This concept was introduced in the form of Barlow Twins for pretraining purposes [Zbon 21]. Another simpler approach is to leverage data augmentation to make the network more aware of these potential variations. To move further in this direction, test time augmentation (TTA) can be used to incorporate multiple reconstruction types as input for a more robust combined prediction. However, a careful cost-benefit analysis must be conducted as each additional parameter variation increases the computational burden. A third approach involves mitigating potential technical bias during training, building upon our previous work on a similar task [Lang 23].

As a side note, this paper also includes an ablation study evaluating the performance of our DL-based approach trained with only 10 % or 20 % of the data. The results of these experiments indicate that achieving the overall good performance of our method requires a substantial amount of training data, especially for the rule-out task, where the data imbalance is more pronounced. This once again underscores two crucial considerations in the field of medical image classification: the limited availability of data often restricts the final performance, and direct comparisons between approaches trained on different data collections are not feasible.

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

How scan parameter choice affects deep learning-based coronary artery disease assessment from computed tomography

Felix Denzinger[1,2]✉, Michael Wels[2], Katharina Breininger[3], Oliver Taubmann[2], Alexander Mühlberg[2], Thomas Allmendinger[2], Mehmet A. Gülsün[2], Max Schöbinger[2], Florian André[4], Sebastian J. Buss[4], Johannes Görich[4], Michael Sühling[2] & Andreas Maier[1]

Recently, algorithms capable of assessing the severity of Coronary Artery Disease (CAD) in form of the Coronary Artery Disease-Reporting and Data System (CAD-RADS) grade from Coronary Computed Tomography Angiography (CCTA) scans using Deep Learning (DL) were proposed. Before considering to apply these algorithms in clinical practice, their robustness regarding different commonly used Computed Tomography (CT)-specific image formation parameters—including denoising strength, slab combination, and reconstruction kernel—needs to be evaluated. For this study, we reconstructed a data set of 500 patient CCTA scans under seven image formation parameter configurations. We select one default configuration and evaluate how varying individual parameters impacts the performance and stability of a typical algorithm for automated CAD assessment from CCTA. This algorithm consists of multiple preprocessing and a DL prediction step. We evaluate the influence of the parameter changes on the entire pipeline and additionally on only the DL step by propagating the centerline extraction results of the default configuration to all others. We consider the standard deviation of the CAD severity prediction grade difference between the default and variation configurations to assess the stability w.r.t. parameter changes. For the full pipeline we observe slight instability (± 0.226 CAD-RADS) for all variations. Predictions are more stable with centerlines propagated from the default to the variation configurations (± 0.122 CAD-RADS), especially for differing denoising strengths (± 0.046 CAD-RADS). However, stacking slabs with sharp boundaries instead of mixing slabs in overlapping regions (called true stack ± 0.313 CAD-RADS) and increasing the sharpness of the reconstruction kernel (± 0.150 CAD-RADS) leads to unstable predictions. Regarding the clinically relevant tasks of excluding CAD (called rule-out; AUC default 0.957, min 0.937) and excluding obstructive CAD (called hold-out; AUC default 0.971, min 0.964) the performance remains on a high level for all variations. Concluding, an influence of reconstruction parameters on the predictions is observed. Especially, scans reconstructed with the true stack parameter need to be treated with caution when using a DL-based method. Also, reconstruction kernels which are underrepresented in the training data increase the prediction uncertainty.

CAD continues to be one of the most severe human diseases with a frequent deadly outcome[1]. Commonly, its root cause is inflammation of perivascular tissue, leading to atherosclerosis, i.e., aggregation of plaque deposits within the vessel walls. These deposits may cause a narrowing of the vessel—so-called stenosis—which may lead to a malperfusion of the heart muscle and therefore cardiac ischemia and a higher risk of acute cardiac death[2]. Also, these plaques can rupture, leading to thrombus formation and thus potentially causing stroke or myocardial infarction. A non-invasive modality capable of assessing the severeness of CAD is CCTA. Contrast agent injected during a Computed Tomography (CT) acquisition enhances the vessels, allowing stenotic lesions to be detected.

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. [2]Computed Tomography, Siemens Healthcare GmbH, Forchheim, Germany. [3]Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. [4]Das Radiologische Zentrum-Radiology Center, Sinsheim-Eberbach-Erbach-Walldorf-Heidelberg, Germany. ✉email: felix.denzinger@fau.de

Commonly, the severeness of CAD, as manifested in CCTA scans, is assessed using the CAD-RADS score[3]). The most severe stenotic lesion within a patient's coronary tree is the main contributor to this score. However, also the location of this culprit lesion and some qualitative aspects are considered when determining this score. Relevant subgroups within the six grades of the CAD-RADS score are CAD-RADS 0, referring to no CAD being present, CAD-RADS 1–2, referring to a non-obstructive CAD without need for further (invasive) assessment and CAD-RADS 3–5 being assigned to patients who should undergo immediate further assessment. The resulting clinical questions are whether a patient has CAD or not (rule-out) and whether a patient has obstructive CAD or not (hold-out). In general, the CAD-RADS score is determined manually by a human reader grading the whole coronary artery tree. This procedure is time-consuming, and with the increasing workload radiologists need to cope with, interest in supporting algorithms is high.

We recently proposed such an algorithm[4] which is Deep Learning (DL)-based and directly predicts the CAD-RADS score using a task-specific data representation and architecture design. A high-level overview of this method is displayed in Fig. 2. It consists of multiple steps: First, the heart is roughly isolated from the rest of the scan[5]. Then, centerlines of the coronary arteries are extracted from the CCTA volume[6] and subdivided into sub-segments. Next, for each of these sub-segments a Multi Planar Reformatted (MPR) volume stack is extracted by interpolating planes orthogonal to each centerline point. Finally, from these MPR volumes, longitudinal views through the centerline are sliced for each respective sub-segment and individually fed into a shared 2D feature extraction Convolutional Neural Network (CNN). The resulting feature representation is used to predict a segment-wise stenosis degree label and global max-pooling of the representations is leveraged to predict the patient-wise CAD-RADS grade and the Agatston score binned according to Rumberger and Kaufman[7] as additional auxiliary target.

This method reaches high performance on the task of regressing all six CAD-RADS grades as well as for the rule-out and hold-out task, with an average accuracy of 0.859 for the six class problem and an AUC of 0.942 and 0.950 for the rule-out and hold-out case, respectively.

Before we go into detail on our methodology in this paper, we want to sketch the bigger picture and discuss variances within the whole measurement system of a CCTA analysis. First, a patient, who exhibits different characteristics like weight, shape, disease state, position, etc., undergoes a CT scan. The resulting projection data is not only influenced by the patient's characteristics but also by the type of scanner, the tube voltage, and the dosage of contrast agent applied. Next, the projection images are reconstructed, whereby the choice of reconstruction kernel, the amount of applied denoising, the heart phase for which the scan is reconstructed, the way neighboring slabs from different heart cycles are stacked together influence the appearance and content of the final volume. Finally, the resulting images are interpreted by a human or Artificial Intelligence (AI) reader. An experienced human reader might be able to disentangle the change in visual perception caused by different acquisition parameters from the actual biological information. However, an AI system, which may have only seen training samples from a subset of fixed scan and reconstruction parameters, is probably influenced by these different technical variations.

Examples for this are already described in literature and can be divided into analyses focusing on the impact of image formation parameter choice on classical Machine Learning (ML)[8–11] on the one hand and on DL[12–14] approaches on the other.

Wielpütz et al.[10] examined the influence of the tube voltage selection and whether Filtered Back Projection (FBP) or Iterative Reconstruction (IR) is used for the volume reconstruction for the task of detecting artificial nodules in an ex vivo study. They found that there was no significant impact on the evaluated classical ML algorithm. In contrast, Berenguer et al.[8], and Li et al.[9] showed that Radiomic features (which include shape-based and first- and second-order statistics on a selected Region of Interest (ROI)) are often not reproducible if one of various scan parameters or the scanner type is varied[8]. Also, the performance of models based on these features may drop[9]. Moreover, Reiazi et al.[11] confirmed that feature distributions vary for different scanner types. For classical ML algorithms, research to compensate differing image formation parameters exist based on statistical assumptions[15] or technical fingerprints in control regions[16].

For DL-based algorithms, analysis of the influence of the image formation parameter was mainly performed on the task of CT lung imaging. Li et al.[12] demonstrated that the performance on the task of detecting nodules changes slightly when the tube voltage or the reconstruction type is varied in a phantom study. A comparable study was conducted by Blazis et al.[13] with a commercially available AI-based system for nodule detection. They used raw data from 24 patients and evaluated 16 different reconstruction settings varying the kernel, denoising strength and reconstruction type. They found an impact of all parameters on the sensitivity of the examined system. Another paper published by Hoang Thi et al.[14] evaluated whether reconstructions with both sharp and soft kernels should be included within training of an algorithm to segment lung nodules. They concluded that the performance is only transferable between kernel types if all options are included in the training step of the algorithm. Recently, the impact of acquisition and patient parameters on an AI-guided CAD assessment system was evaluated[17]. The underlying pipeline consists of ML-based centerline extraction and labeling, inner and outer wall segmentation and lesion detection and scoring systems. However, the final prediction of each step is double-checked by a human reader to prevent error propagation. They explore several different variations of acquisition and patient parameters including the scanner type, tube voltage, gating technique, several clinical parameters, etc.. Limitations of this work are that the individual subgroups differ in size and that the impact of a single parameter change on the system cannot be directly measured but needs to be statistically assessed over a large patient population. Furthermore, the influence of the variations on the AI components cannot be separated from the additional human reader.

After this brief overview of related work, we will define the scope of this work. The image formation parameter choice often differs for different clinical sites and personal preference. Therefore, when considering the clinical application of an AI-based CAD approach, it is crucial to evaluate how well this algorithm performs for

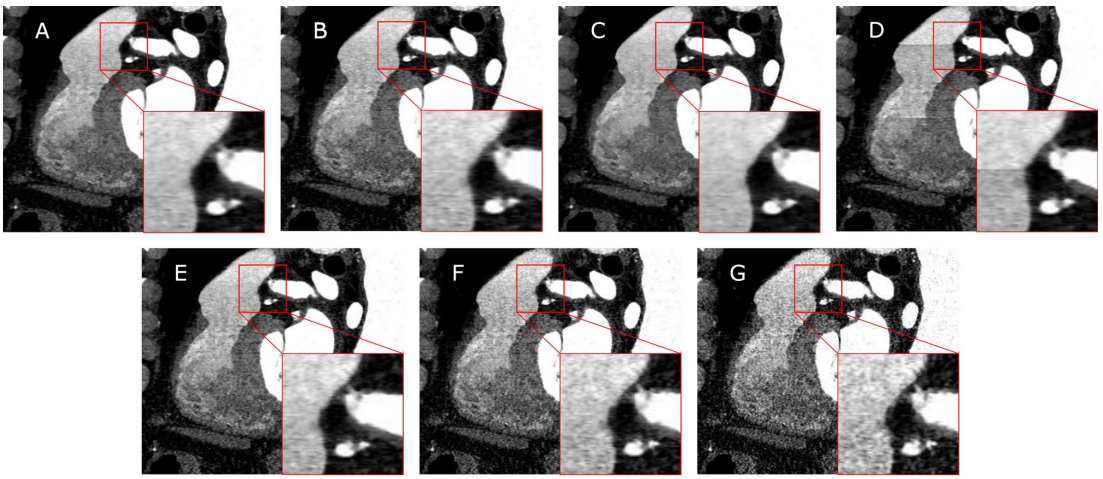

Figure 1. Sagittal views of a CCTA scan reconstructed with different parameters. The default configuration (**A**) reconstructed with an Advanced Modeled Iterative Reconstruction (ADMIRE) strength of 3, mixed stack and Bv36 kernel is varied by using: an ADMIRE strength of 2 (**B**) or 4 (**C**), using true stack (**D**) and utilizing a Bv40 (**E**), Bv44 (**F**) or Bv49 (**G**) reconstruction kernel.

differing reconstruction parameters. With this work, we aim to analyse the influence of a set of reconstruction parameters on our previously published CAD-RADS scoring system[4]. These parameters are defined in more detail in "Raw prediction changes" and include the ADMIRE strength, stacking and reconstruction kernel choice. Image impressions of these altered reconstruction parameters are depicted in Fig. 1. To systematically evaluate the influence of each of these parameters on our AI system, we leverage a collection of 500 raw data sets, and reconstruct all samples with one default configuration and single parameter variations. The CAD-RADS grading method consists of several preprocessing and a Neural Network (NN) prediction step. To be able to separate the variation changes' impact on the NN step of the pipeline from the impact on the preprocessing, we evaluate the full pipeline and the pipeline with the centerlines propagated from the default configuration to all variations. Our contributions can be summarized as follows:

- To the best of our knowledge, we conduct the first evaluation of scan parameter dependency of a DL-based approach for automatic assessment of CCTA scans with paired data, i.e. the sole difference between the individual reconstructions being the parameter change.
- We separate the influence of the parameter changes on the preprocessing results from the change in image data.
- We provide guidance regarding which image formation parameters need to be treated with caution.

Methods

All the methods in this study were performed in accordance with the Declaration of Helsinki.

Data. Before going into detail about the data used in this study, the distribution of labels and reconstruction parameters, we want to define the parameter space we evaluate. An exemplary case for each parameter configuration is displayed in Fig. 1.

ADMIRE strength. A parameter that might influence the prediction and is sometimes altered in clinical practice is the number of iterations of the reconstruction algorithm. In this study, we use the ADMIRE algorithm[18] to denoise already during reconstruction. Depending on the number of iterations, the algorithm reduces noise but also may introduce denoising artifacts. A popular choice in clinical practice is a ADMIRE strength of 3 (default) with variations to 2 or 4 depending on the image quality and personal reader preference. We therefore reconstructed our raw data with these three parameter choices.

Stacking. Mostly, CCTA projection images are acquired over multiple heart cycles as the field of view of the detector is usually not large enough to cover the whole heart in one rotation. Therefore, the patient table is moved along the superior-inferior axis during the acquisition. In addition, depending on the motion occurring between heart cycles, e.g. breathing motion, the projections for different z-positions may not match each other directly at their boundaries. Since there are usually overlapping regions for the patient positions, there are two possible strategies to cope with this: either the overlapping regions are merged using interpolation strategies (mixed stack, default), or only the information of a single position is preserved (true stack). As the first strategy may introduce artifacts when a lot of motion occurs between the heart cycles, physicians prefer the true stack

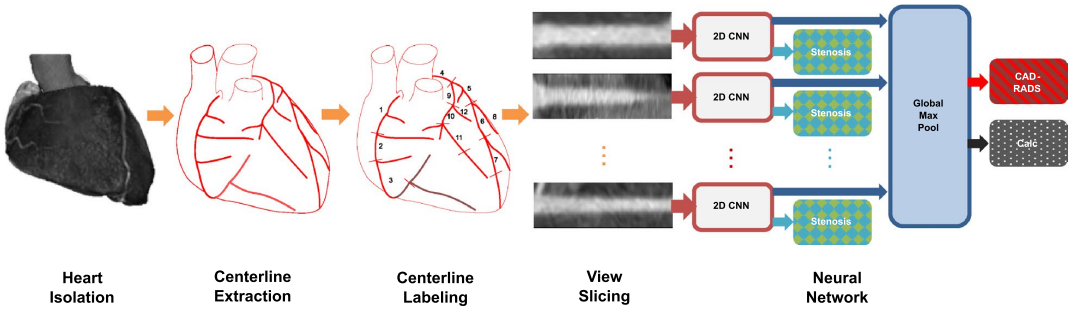

Figure 2. Overview of the used AI approach. First the heart is isolated from the CCTA scan using the algorithm proposed by Zheng et al.[5]. Then the coronary centerlines are extracted[6]. These are subdivided into up to 11 equally sized sub-segments[4]. For each sub-segment, longitudinal slices are interpolated orthogonal to the centerline and fed into a NN.

option in these cases. However, it leads to sharp boundaries between the individual stacks, which do not necessarily impact human readers but may impact the performance of algorithms processing the volume (cf. Fig. 4).

Reconstruction kernel. Finally, another important parameter that is often changed in clinical practice is the reconstruction kernel. By adapting it, the sharpness of the edges can be increased at the cost of an increased noise level. Each vendor offers its own set of reconstruction kernels. For our experiments, we choose the Siemens Healthineers specific Bv36 kernel as the default, which is a medium sharp kernel specifically designed for the heart anatomy and therefore commonly used for CCTA scans. As we observed an increase in instability correlating with kernel sharpness during initial experiments we chose the increasingly sharp Bv40, Bv44, and Bv49 kernels as variations. With all of these variations, the volume content should be consistent while the appearance may change (cf. Fig. 1).

Data characteristics. We use two data collections in this study. Once, a data collection of 2596 reconstructed CT scans (data set A) as training set for the CAD-RADS scoring system. Additionally, we leverage a data collection containing raw CT data of 500 patients (data set B). Both data collections were collected at the same center with Siemens SOMATOM Force scanners. All samples in data collection A were reconstructed using the Bv36 reconstruction kernel with a slice thickness of 0.6 mm. Furthermore, the ADMIRE reconstruction algorithm was applied with a strength of 3. 55 cases were reconstructed using true stack and all others with mixed stacking. The CAD-RADS class frequency in the training set (A) is 370, 551, 828, 542, 281, 24 for CAD-RADS 0 to 5 respectively. For the raw data collection B 7 configurations (examples displayed in Fig. 1) were reconstructed for all 500 data samples: a default configuration (ADMIRE strength = 3; stacking = mixed; kernel = Bv36) varied by using an ADMIRE strength of 2 or 4, true stacking and a Bv40, Bv44 or Bv49 reconstruction kernel. Reconstruction was performed with ReconCT (version 15.0, Siemens Healthineers). For data set B the class distribution is more balanced with 73, 61, 81, 85, 146, 54 samples for each respective CAD-RADS grade.

Algorithm. A high-level overview of the method evaluated in this work is depicted in Fig. 2. As this scientific publication focuses on the evaluation of scan parameter influences, we refer to the publication where the evaluated method was proposed[4] for most details. Still, we want to mention some properties which have impact on the robustness analysis. From the CCTA scan data to the final prediction, multiple different algorithms are utilized. These include an algorithm for creating a rough segmentation of the heart[5], extracting the centerlines[6], and labeling them[4]. As each of the later steps depends on the preceding ones, differences are propagated through the whole pipeline, altering the final prediction. Centerline labeling does not depend on the image data but solely leverages the centerline coordinates. The last pipeline step is the data processing through a task-specific DL-based architecture. One forward pass of this architecture takes one longitudinal slice of each labeled coronary segment as an input. In order to include all information in the final prediction the concept of Test Time Augmentation (TTA) is leveraged by extracting these longitudinal slices at 16 equidistant angles around the centerline as rotation axis within a range of $[0, \pi]$ from the volume. The average prediction over all angles is considered the final prediction for a single model. This is done to prevent the algorithm from missing information due to unfortunate angle choices. Also, the models of 25 training runs with different randomly chosen training and validation splits are ensembled to increase the method's stability and performance. We encode the prediction in a multi-label format. As, if a patient belongs, e.g., to the CAD-RADS 3 category, he also fulfills the criteria of CAD-RADS 0-2 due to a gradual nature of the score (i.e., (1, 1, 1, 1, 0, 0) represents CAD-RADS 3). We consider the predicted cumulative probability of all classes as the raw output score. Due to class imbalance default thresholds of .5 between the raw predictions do not necessarily lead to optimal class predictions. To circumvent this problem we determine more optimal thresholds: we define them as the percentiles of the raw prediction histogram. The percentile values are defined by the class frequencies. E.g. if 73 cases belong to the CAD-RADS 0 and 61 to the CAD-RADS 1 class, the threshold between CAD-RADS 1 and CAD-RADS 2 is the 134th lowest prediction.

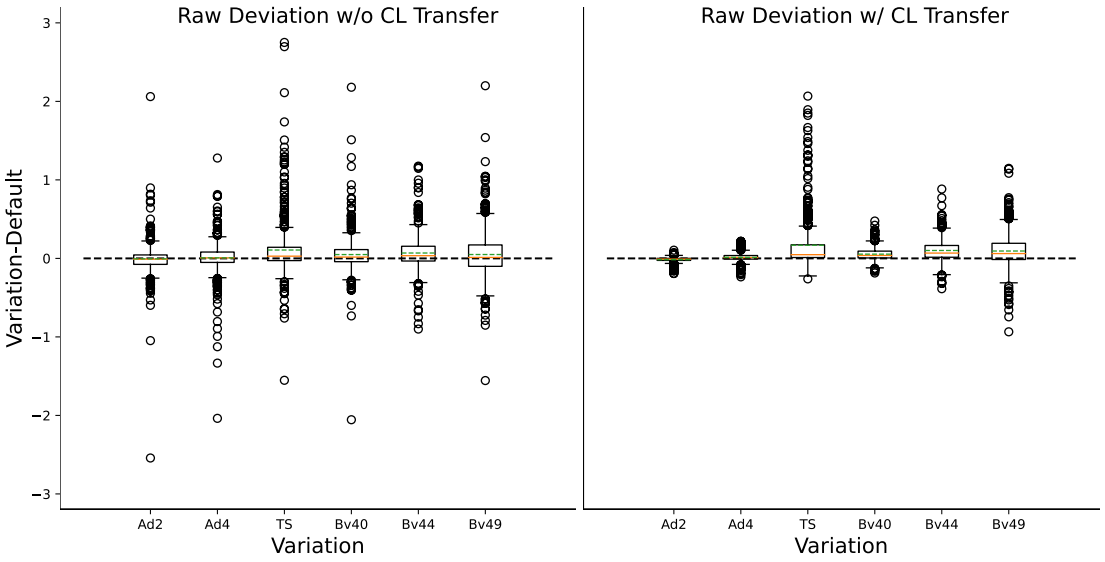

Figure 3. Boxplot of the rescaled raw prediction difference between the default configuration and all variation configurations (Ad2, Ad4 = ADMIRE strength of 2/4; TS = true stack; Bv40, Bv44, Bv49 = different reconstruction kernels), with and without centerline propagation.

As we focus on having the best possible class separation for the robustness analysis, we calculated the thresholds on the test set predictions of the default configuration reconstructions. For the performance analysis we use the threshold invariant AUC metric.

Evaluation. As the focus of this work is the evaluation of the stability of our CAD-RADS estimation approach, we choose to compare predictions of the default parameter configuration to the individual parameter variations. Since with the step of binning the network's CAD-RADS predictions to distinct scores a lot of information is lost we choose to evaluate the difference in raw prediction scores. These also encode a kind of certainty regarding the prediction. To render the raw scores comparable for all individual CAD-RADS grades we rescale the predictions such that the value range between two thresholds always equals 1. As metrics, we evaluate whether the parameter change leads to a shift in the mean prediction and how much the standard deviation over all patients changed. Also, the number of cases where the parameter change leads to a different binned prediction is of interest, although outliers may influence it. Finally, the overall performance of the method regarding the hold-out and rule-out case may vary. Here, we decide to focus on the AUC as a threshold independent metric, also because the thresholds were defined on the test set. Also, to separate the influence of the parameter change on the NN component; we evaluate the influence of the deviations if we propagate the centerlines extracted from the default configuration to all others. Furthermore, as our approach relies on model ensembling, TTA, and a large training data cohort, which are all known factors to increase the robustness of DL-based models, we conduct additional experiments without model ensembling, without TTA, and with random subsets of only 10% (259 patients) or 20% (519 patients) of the training data.

Ethical standards. The CT examinations were clinically indicated by the referring physicians and conducted in accordance with current clinical standards, guidelines, and recommendations. The study was performed in accordance with the Declaration of Helsinki and was approved by the local ethics committee (S-226/2016 and S-758/2018, Ethikkommission der Medizinischen Fakultät Heidelberg, Germany). Subjects included as of January 2019 gave informed consent in the scientific data analyses. For the retrospective analyses of the datasets acquired before January 2019, a waiver of consent was granted by the aforementioned ethics committee.

Results

Raw prediction changes. First, we want to report the changes in raw predictions caused by varying the image formation parameters. Therefore, the prediction difference between the default and the respective variation for the whole data set is visualized as boxplot in Fig. 3. The standard deviation of the distributions displayed in Fig. 3 can be seen in Table 1. From a first glance, it is apparent that propagating the centerlines from the default to the varied configuration leads to a decreased variance. This holds true for all variations when comparing the standard deviations. For a differing amount of denoising iterations of the ADMIRE algorithm, the variance is relatively low in the case centerlines are propagated. For sharper kernels, mean offsets are observed ($\mu_{Bv40} = 0.054$; $\mu_{Bv44} = 0.101$; $\mu_{Bv49} = 0.094$). The most considerable offset of the mean value with an amplitude of 0.167 is observed for the true stack variation with centerline propagation. When using true stack the standard devia-

	CL transfer	Ad2	Ad4	TS	Bv40	Bv44	Bv49	Mean
Default	Without	0.220	0.225	0.355	0.243	0.233	0.307	0.226
No ensembling	Without	0.270	0.280	0.397	0.292	0.309	0.418	0.281
No TTA	Without	0.267	0.280	0.398	0.286	0.305	0.392	0.275
10% of data	Without	0.405	0.411	0.462	0.413	0.468	0.569	0.390
20% of data	Without	0.297	0.318	0.360	0.304	0.340	0.464	0.298
Default	With	0.035	0.057	0.313	0.080	0.143	0.227	0.122
No ensembling	With	0.048	0.087	0.340	0.110	0.199	0.334	0.160
No TTA	With	0.043	0.075	0.347	0.108	0.192	0.296	0.152
10% of data	With	0.030	0.077	0.286	0.148	0.261	0.433	0.176
20% of data	With	0.051	0.077	0.308	0.106	0.209	0.376	0.161

Table 1. Standard deviation of the raw prediction change for all individual variations (abbreviations as in Fig. 3) compared to the default. "Centerline (CL) Transfer" refers to the centerlines being propagated from the default to the varied configurations.

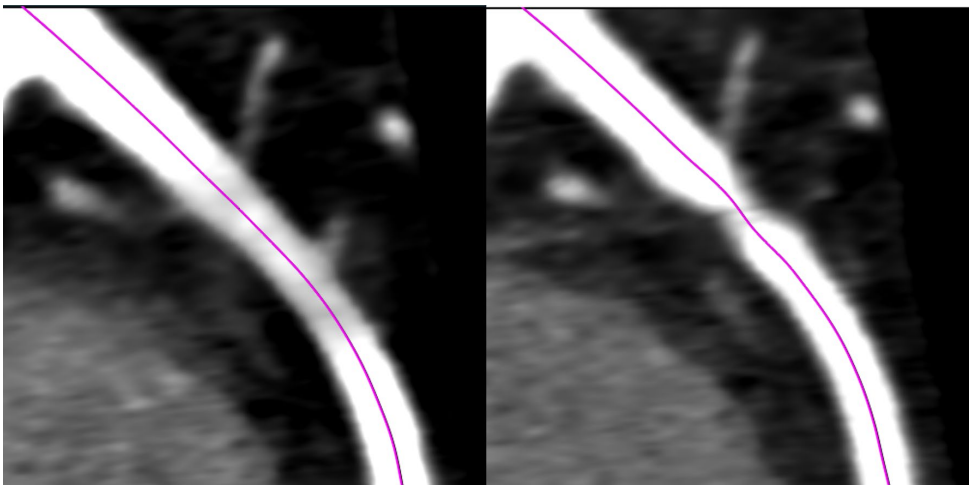

Figure 4. Curved Planar Reformatted (CPR) view of the Right Coronary Ascending (RCA) proximal segment for a reconstruction with mixed stacking (left) and true stacking (right) of the same raw data set. Due to the sharp slab boundaries the visual perception suggests a narrowing of the vessel.

tion and therefore the variance of the prediction change is higher than with all other variations. An explanation for this behavior is that vessels at the slab edges may have the visual impression of being narrowed due to the sharp slab boundaries. A visual example of this effect is shown in Fig. 4. Analyzing different reconstruction kernels, the resulting variance in prediction seems to correlate with the sharpness of the kernel.

Above mentioned trends observed for the variations hold when assessing the model without ensembling, TTA, or using less training data (cf. Table 1). However, the general robustness decreases for each of these experiments compared to the default model. The highest standard deviation (± 0.390) of the prediction changes is observed when using just 10 % of the data without propagating the centerline extraction results. Results are again more robust when propagating the centerlines. However, the standard deviation still increases by at least 32% when not using TTA and up to 44% when only using 10% of the data.

Binned prediction changes. To more directly assess the impact on the resulting clinical scores and decisions, we also show how many times the prediction changed due to the changed image formation parameters. We therefore present the number of class changes in Fig. 5. Overall, for all configuration and the full pipeline between 12 and 20% of the cases changed the predicted CAD-RADS score. A low number changes between the clinically relevant cases of rule-out and hold-out. Moreover, the same trends as described for the other metric hold true for all varied configurations.

Appearance changes. To foster intuition on why the reconstruction parameter changes lead to different predictions, we depict the stretched proximal RCA segment as fed into the NN for the cases with the respective largest CAD-RADS prediction change for each variation in Fig. 6 (all other segments are provided in the

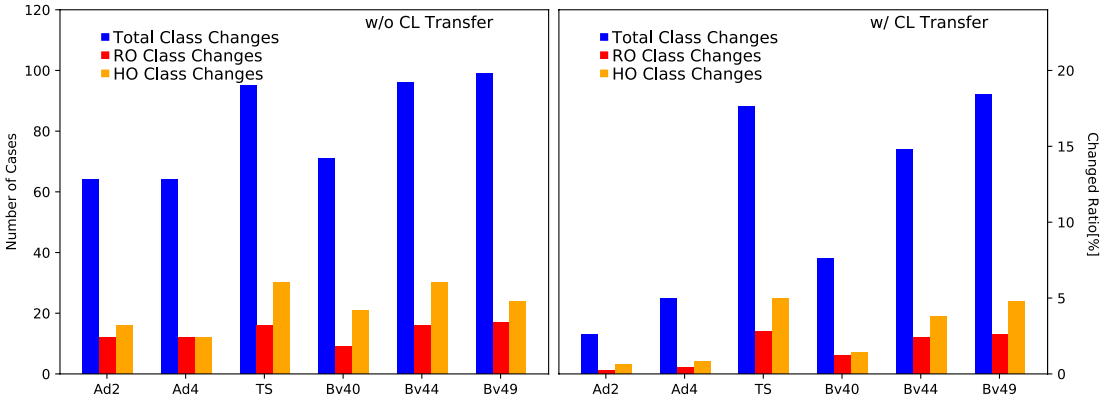

Figure 5. Number of samples for which the class prediction changed due to the parameter change (abbreviations as in Fig. 3). Total class changes refers to all CAD-RADS grades, RO to the rule-out and HO to the hold-out task. Note: predictions close to the thresholds easily change class bin even with small prediction changes.

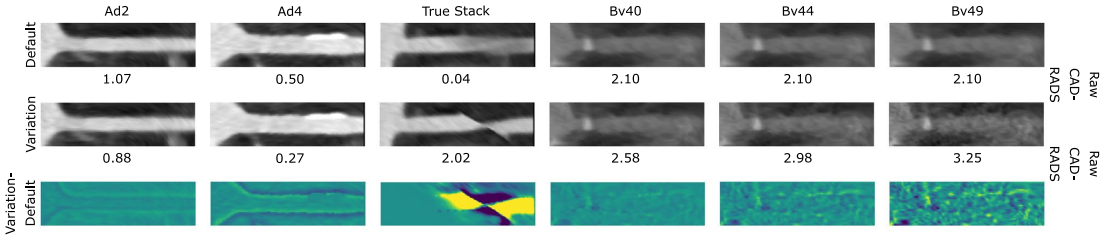

Figure 6. Proximal RCA segments for the cases with the largest CAD-RADS prediction deviation for each parameter configuration (note that the same patient showed the largest deviation for all possible kernel choices). Raw predictions with propagated preprocessing results are displayed for each respective configuration.

Supplementary material). When changing the ADMIRE strength to 2 or 4 the visual appearance is quite similar for a human reader, but the difference image shows slight deviations, especially around the vessel wall, which may explain the slightly different scores. For the true stack variations, the reason for the differing prediction is already apparent when looking at the deviation image: the slab boundary cuts through the vessel obscuring the image information. Lastly, we assess how the reconstruction kernel choice changes the appearance. Here we can see incrementally higher noise levels which appear slightly localized at the vessel wall as seen in the difference images.

Overall performance. Besides individual prediction changes, the method's overall performance is of interest. For the default configuration an AUC of 0.957 (95% CI [0.942, 0.971]) for the rule-out task and 0.971 (95% CI [0.961, 0.981]) for the hold-out task is achieved as displayed in Fig. 7 and in Table 2. The deviation from the results reported in Denzinger et al.[4] is caused by the different class balance/test set evaluated. In Fig. 7 the performance deviation for the different variations is displayed as well. Interestingly, the deviation is mostly within the CI and, therefore, insignificant in these cases. The only variation leading to a significant performance drop is the use of true stack instead of mixed stacking, but only for the rule-out task. A possible explanation for this is that vessels at the stack boundaries may appear stenotic due to the sudden jump between stacks (see Fig. 4). Above observations also hold true when the centerlines are propagated from the default to the variation configurations.

Assessing the performance changes for different model configurations (no ensembling, no TTA, less training data), the findings for the default configuration hold when using no ensembling and no TTA. When training with only 10% or 20%, we can observe a significant drop in performance of our approach with larger confidence intervals, especially on the rule-out task, indicating a model that did not generalize as well. Interestingly, the performance actually improves for some of the variations. This indicates that the model focuses on different features when trained with less data and that these features are actually enhanced when the noise level varies compared to the default configuration. This reasoning at least applies to the rule-out task, where the task inherent class imbalance impacts the generalization of the model more when reducing the amount of data. Overall, the performance changes are still mostly within the 95% CIs.

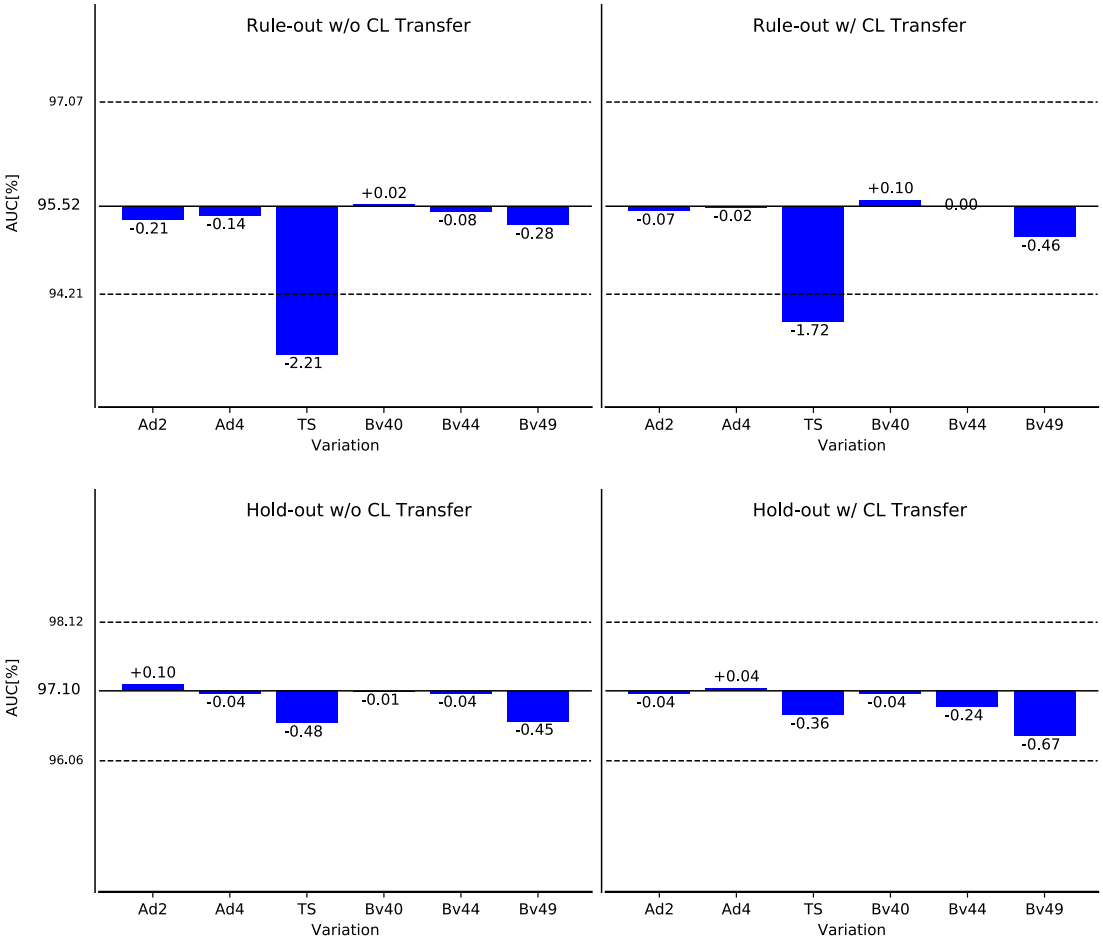

Figure 7. Performance on the data set A with all configurations (abbreviations as in Table 1) compared to the default for the rule-out and hold-out task, with and without centerline propagation. The dashed lines correspond to the 95% Confidence Interval (CI) for the default configuration. Note that the performance of the default configuration does not depend on the preprocessing as the centerlines of the default configuration are propagated.

Discussion

For all evaluation metrics, it becomes apparent that the preprocessing steps have an impact on the prediction if the scan parameters are varied. The centerline extraction is likely a larger contributor regarding this behaviour as small changes in the heart isolation mask are not expected to lead to much of a difference and the centerline labeling approach only depends on the centerline points. A detailed evaluation regarding the influence of scan parameters on the centerline extraction results is not the main focus of this work, but should be subject of further research. That said, looking at the overall performance of the method, there is mostly no significant performance drop, also when considering the full pipeline. A possible explanation for this behavior is that a similar number of cases are correct for any one variation as were previously erroneous as no parameter variation causes a mean shift. However, there are two perspectives (global vs. local) regarding performance, and knowing that a slight parameter change may lead to a different diagnosis by the system for a single patient does not build trust. On the other hand, when comparing algorithms with the current gold standard—manual assessment by physicians—one must acknowledge that different readers (or one reader over time) may also grade the same or different reconstructions differently. In literature, the inter-observer variability of manual CAD-RADS scoring is reported with an inter-observer correlation (IOC) of 0.748 (average pairwise inter-observer agreement (IOA) 0.847)[19], an IOA of 0.885[20], or an IOC of 0.958[21], depending on the study design and reader experience. When considering the ratio of unchanged predictions (cf. Fig. 5, right) as a metric comparable to the IOA, varying parameters like the denoising strength, and a slightly sharper reconstruction kernel are within this range. However, for the true stack configuration and even sharper kernels, the number of changed predictions increases. Looking at Fig. 1 and Fig. 6, these variations have the largest impact on the visual perception of human readers as well and

	Task	CL transfer	AUC [%]	Ad2	Ad4	TS	Bv40	Bv44	Bv49
Default	Rule-out	Without	95.72 [94.21, 97.07]	− 0.14	− 0.15	− 2.07	− 0.08	− 0.08	− 0.36
No ensembling	Rule-out	Without	94.65 ± 0.74	+ 0.03	− 0.15	− 1.98	+ 0.05	− 0.06	− 0.23
No TTA	Rule-out	Without	94.59 ± 0.39	− 0.09	− 0.04	− 1.58	+ 0.04	− 0.15	− 0.68
10% of Data	Rule-out	Without	86.73 [84.01, 89.52]	+ 1.25	+ 1.97	− 0.40	+ 1.29	+ 1.95	+ 2.41
20% of Data	Rule-out	Without	87.84 [85.00, 90.43]	+ 1.07	+ 0.99	− 0.69	+ 0.78	+ 1.07	+ 1.15
Default	Hold-out	Without	97.11 [96.06, 98.12]	+ 0.08	− 0.04	− 0.34	− 0.03	− 0.06	− 0.48
No ensembling	Hold-out	Without	96.79 ± 0.18	+ 0.12	− 0.13	− 0.44	− 0.06	− 0.16	− 0.62
No TTA	Hold-out	Without	96.66 ± 0.32	− 0.09	+ 0.08	− 0.42	− 0.08	− 0.30	− 0.88
10% of data	Hold-out	Without	93.01 [90.97, 94.86]	+ 0.39	+ 0.20	+ 0.08	+ 0.14	− 0.14	− 0.06
20% of data	Hold-out	Without	96.00 [94.75, 97.20]	− 0.19	+ 0.09	− 0.57	− 0.35	− 0.79	− 1.01
Default	Rule-out	With	95.72 [94.21, 97.07]	− 0.06	− 0.04	− 1.66	+ 0.06	− 0.06	− 0.54
No ensembling	Rule-out	With	94.65 ± 0.74	− 0.05	− 0.13	− 1.42	+ 0.12	+ 0.02	− 0.51
No TTA	Rule-out	With	94.59 ± 0.39	− 0.12	− 0.12	− 2.12	− 0.11	− 0.10	− 0.42
10% of data	Rule-out	With	86.73 [84.01, 89.52]	+ 0.05	+ 0.07	− 0.59	+ 0.75	+ 1.02	+ 1.24
20% of data	Rule-out	With	87.84 [85.00, 90.43]	+ 0.13	− 0.27	− 1.48	+ 0.52	+ 1.10	+ 0.92
Default	Hold-out	With	97.11 [96.06, 98.12]	− 0.05	+ 0.05	− 0.20	− 0.05	− 0.25	− 0.71
No ensembling	Hold-out	With	96.79 ± 0.18	− 0.04	+ 0.02	− 0.36	− 0.10	− 0.30	− 0.83
No TTA	Hold-out	With	96.66 ± 0.32	+ 0.07	− 0.10	− 0.48	− 0.08	− 0.17	− 0.70
10% of data	Hold-out	With	93.01 [90.97, 94.86]	+ 0.10	− 0.18	− 0.08	+ 0.07	+ 0.11	− 0.20
20% of data	Hold-out	With	96.00 [94.75, 97.20]	− 0.07	+ 0.10	− 0.83	− 0.12	− 0.33	− 0.88

Table 2. Performance deviation for all individual variations (abbreviations as in Fig. 3) compared to the default. "CL Transfer" refers to the centerlines being propagated from the default to the varied configurations. AUC is either displayed with 95% CI or the standard deviation over all respective single models or angles.

may even change the perception regarding the disease state as can be seen in Fig. 4. Also, reconstruction with a sharper kernel might lead to such a high noise level that the resulting volumes are hard to read. However, no study comparing the performance of readers on the task of CAD-RADS grading for differing reconstruction kernels exist. Such a study would be hard to design as readers might be biased by their first reading under a different reconstruction configuration.

Therefore, the algorithm's variance in prediction appears to be within the range of human readers. However, this is usually not the motivation to use algorithms as assistance tools. Algorithms are expected to yield consistent outcomes for the same patient. In particular, since the used algorithm embodiment mostly behaves as a black box, a higher robustness with respect to parameter variation is required to allow for clinical acceptance. A possible way to achieve this robustness is to include different parameter configurations into training[14]. This idea seems promising, as in the current training pipeline only reconstructions with a softer kernel are included. An additional possibility to increase robustness might be to transfer algorithms aiming to disentangle biological and technical information[16] into the deep learning world.

Another aspect to elaborate on is whether the results reported here are transferable to other methods proposed in literature[22–30]. These mostly focus on the determination and detection of significant stenosis which is similar to the hold-out task. Usually, these approaches also rely on a prior centerline extraction usually followed by an MPR volume construction[22–27,29]. Additionally, for all approaches a CNN is used as a feature extractor. Therefore, findings presented here should be largely transferable to different architecture embodiments. Exceptions may be the works of Muscoguiri et al.[28], who directly operate on the 3D data, and Paul et al.[30], who operate on the curved views instead. Furthermore, all of the above-mentioned approaches were trained from data collected from a respective single site. As there usually is an internal consensus on how data is reconstructed at each individual clinical site, our choice of training our method with the data reconstructed as part of the clinical workflow is a valid and transferable choice.

Also, we evaluate whether our CAD-RADS scoring NN behaves similarly if no robustness enhancing measures like ensembling and TTA, or a smaller data collection are used. We have shown that model ensembling and TTA did not alter the findings of our study. However, a limited amount of training data leads to a less generalized model, especially regarding the rule-out task, which is most severely impacted by the task inherent class imbalance. Related works usually perform analysis on a per-vessel basis and are therefore not impacted as severely by this class imbalance. Still, the uncertainty when considering the standard deviation of the raw CAD-RADS prediction changes as a metric, behaves comparable to the default configuration.

Conclusion

In this work, we analyzed the effects of varying image formation parameters on an existing AI-based system to automatically grade CCTA scans with the CAD-RADS score. To this end, we reconstructed 500 raw CCTA scans under eight parameter configurations, which to our knowledge are commonly applied in clinical practice. Parameter changes evaluated include the denoising strength, slab combination, and reconstruction kernel choice.

We found that the preprocessing steps as well as the NN prediction step are not robust to all parameter variations. Using true stack to combine slabs of different heart phases leads to a slight overestimation of the CAD-RADS score for patients with movement between slabs as stack artifacts occurred. These artifacts can create the visual perception of a narrowed vessel at slab boundaries. We conclude that one should consider excluding datasets reconstructed with this parameter from training and application. For varied reconstruction kernels, the variance of the prediction change increased with increasing kernel sharpness. Globally, the performance remained on a high level for all variations. However, individual prediction changes occurred, which may not built trust in clinical application of such an algorithm if a patient's scoring depends on the way their scan was reconstructed. Therefore, we conclude that strategies to create more robust predictions for individual patients need to be developed. These may include the use of a more diverse training set. However, also the preprocessing steps need some additional attention as they were contributors to the prediction changes. We have shown that the same findings hold true when leaving out robustness-enhancing measures like model ensembling and TTA. Furthermore, the method at hand behaves slightly differently when trained with less samples due to reduced generalization.

Data availability

The data are not publicly available due to data protection regulations. They are available from the authors upon reasonable request.

Received: 16 August 2022; Accepted: 2 February 2023
Published online: 13 February 2023

References

1. Roth, G. *et al.* Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study. *J. Am. Coll. Cardiol.* **76**, 2982–3021 (2020).
2. Fuster, V., Badimon, L., Badimon, J. J. & Chesebro, J. H. The pathogenesis of coronary artery disease and the acute coronary syndromes. *N. Engl. J. Med.* **326**, 310–318 (1992).
3. Cury, R. C. *et al.* Coronary artery disease-reporting and data system (CAD-RADS): An expert consensus document of SCCT, ACR and NASCI: Endorsed by the ACC. *JACC CI* **9**, 1099–1113 (2016).
4. Denzinger, F. *et al. CAD-RADS Scoring Using Deep Learning and Task-Specific Centerline Labeling*. (MIDL, 2022).
5. Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M. & Comaniciu, D. Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features. *IEEE Trans. Med. Imaging* **27**, 1668–1681 (2008).
6. Zheng, Y., Tek, H. & Funka-Lea, G. Robust and accurate coronary artery centerline extraction in CTA by combining model-driven and data-driven approaches. In International Conference on *MICCAI* 74–81 (Springer, 2013).
7. Rumberger, J. & Kaufman, L. A rosetta stone for coronary calcium risk stratification: Agatston, volume, and mass scores in 11,490 individuals. *Am. J. Roentgenol.* **181**, 743–748 (2003).
8. Berenguer, R. *et al.* Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters. *Radiology* **288**, 407–415 (2018).
9. Li, Y. *et al.* CT slice thickness and convolution kernel affect performance of a radiomic model for predicting EGFR status in non-small cell lung cancer: A preliminary study. *Sci. Rep.* **8**, 1–10 (2018).
10. Wielpütz, M. O. *et al.* Computer-aided detection of artificial pulmonary nodules using an ex vivo lung phantom: Influence of exposure parameters and iterative reconstruction. *Eur. J. Radiol.* **84**, 1005–1011 (2015).
11. Reiazi, R. *et al.* Prediction of human papillomavirus (HPV) association of oropharyngeal cancer (OPC) using radiomics: The impact of the variation of CT scanner. *Cancers* **13**, 2269 (2021).
12. Li, Q. *et al.* Detectability of pulmonary nodules by deep learning: Results from a phantom study. *Chin. J. Acad. Radiol.* **2**, 1–12 (2019).
13. Blazis, S. P., Dickerscheid, D. B., Linsen, P. V. & Jarnalo, C. O. M. Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system. *Eur. J. Radiol.* **136**, 109526 (2021).
14. Hoang-Thi, T.-N. *et al.* Deep learning for lung disease segmentation on CT: Which reconstruction kernel should be used?. *Diagn. Interv. Imaging* **102**, 691–695. https://doi.org/10.1016/j.diii.2021.10.001 (2021).
15. Ibrahim, A. *et al.* The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without combat harmonization. *Cancers* **13**, 1848 (2021).
16. Mühlberg, A. *et al.* The technome-a predictive internal calibration approach for quantitative imaging biomarker research. *Sci. Rep.* **10**, 1–15 (2020).
17. Jonas, R. A. *et al.* The effect of scan and patient parameters on the diagnostic performance of AI for detecting coronary stenosis on coronary CT angiography. *Clin. Imaging*https://doi.org/10.1016/j.clinimag.2022.01.016 *(2022).*
18. Ramirez-Giraldo, J. C., Grant, K. L. & Raupach, R. *ADMIRE: Advanced Modeled Iterative Reconstruction*. (Siemens Healthc., 2018).
19. Hu, J.-Y. *et al.* Interobserver reliability of the coronary artery disease reporting and data system in clinical practice. *J. Thorac. Imaging* **36**, 95–101 (2021).
20. Razek, A. A. K. A. *et al.* Inter-observer agreement of the coronary artery disease reporting and data system (CAD-RADSTM) in patients with stable chest pain. *Pol. J. Radiol.* **83**, e151 (2018).
21. Maroules, C. D. *et al.* Coronary artery disease reporting and data system (CAD-RADSTM): Inter-observer agreement for assessment categories and modifiers. *J. Cardiovasc. Comput. Tomogr.* **12**, 125–130 (2018).
22. Zreik, M. *et al.* A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography. *IEEE Trans. Med. Imaging* **38**, 1588–1598 (2018).
23. Candemir, S. *et al.* Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network. *Comput. Med. Imaging Graph.* **83**, 101721 (2020).
24. Denzinger, F. *et al.* Coronary artery plaque characterization from CCTA scans using deep learning and radiomics. In *International Conference on MICCAI* 593–601 (Springer, 2019).
25. Tejero-de-Pablos, A. *et al.* Texture-based classification of significant stenosis in CCTA multi-view images of coronary arteries. In International Conference on *MICCAI* 732–740 (Springer, 2019).
26. Denzinger, F. *et al.* Deep learning algorithms for coronary artery plaque characterisation from CCTA scans. In *BVM, LNCS 2020*, 193–198 (Springer, 2020).
27. Denzinger, F. *et al.* Automatic CAD-RADS scoring using deep learning. In *International Conference on MICCAI*, 45–54 (Springer, 2020).
28. Muscogiuri, G. *et al.* Performance of a deep learning algorithm for the evaluation of CAD-RADS classification with CCTA. *Atherosclerosis* **294**, 25–32 (2020).

29. Ma, X., Luo, G., Wang, W. & Wang, K. Transformer network for significant stenosis detection in CCTA of coronary arteries. In *International Conference on MICCAI*, 516–525 (Springer, 2021).
30. Paul, J.-F., Rohnean, A., Giroussens, H., Pressat-Laffouilhere, T. & Wong, T. Evaluation of a deep learning model on coronary CT angiography for automatic stenosis detection. *Diagn. Interv. Imaging* (2022).

Acknowledgements
This work was partially funded by Siemens Healtcare GmbH, Erlangen, Germany. K.B. gratefully acknowledges the support of the project “Dhip campus-bavarian aim”. We acknowledge financial support by Deutsche Forschungsgemeinschaft and Friedrich-Alexander-Universität Erlangen-Nürnberg within the funding programme "Open Access Publication Funding".

Disclaimer
The methods and information presented here are based on research and are not commercially available.

Author contributions
F.D., M.W., K.B. and O.T. conceived the experiments. F.D. conducted the experiments and analyzed the results. T.A. and M.Sc. provided technical background. M.A.G. provided supporting algorithms. F.A., S.J.B. and J.G. provided the data and medical background. M.Su. and A.Ma. supervised the study. The manuscript was written by F.D., M.W., K.B., O.T. and A.Mu. All authors reviewed the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
F.D., M.W., O.T., A.Mu., M.A.G., M.Sc. and M.Su. are employees of Siemens Healthcare GmbH. F.A., S.J.B. and J.G. received a research grant from Siemens Healthcare GmbH. All other authors do not have any other conflicts to declare.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-29347-9.

Correspondence and requests for materials should be addressed to F.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

7.2 Handling Label Uncertainty and Shepherd's Crook RCA Detection

The final research topic to be discussed within this thesis was published as:

[Denz 23b] F. Denzinger, M. Wels, O. Taubmann, F. Kordon, F. Wagner, S. Mehltretter, M. A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Sühling, A. Maier, and K. Breininger. "Handling Label Uncertainty on the Example of Automatic Detection of Shepherd's Crook RCA in Coronary CT Angiography". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023

Within it, we contributed the following aspects to the field of research:

- Providing the first approach to automatically determine the Shepherd's crook (SC) right coronary artery (RCA) from centerline data.
- Evaluating a set of strategies to handle non-confidently labeled samples during training time.
- Proposing a straight forward quantile-based abstention rule.

7.2.1 Publication Overview

Let's begin by discussing the clinical motivation behind this research. As explained in Section 1.1.3, the SC RCA is a norm variant characterized by a highly tortuous turn immediately after the ostium. It is important to automatically detect the presence of SC RCA in screening scans, as it complicates minimally invasive procedures. To address this task, we curated a dataset of 519 cases and annotated them with respect to SC RCA presence. However, during the annotation process, some samples could not be confidently labeled as exhibiting SC RCA or not, leading us to designate a subset of samples as unsure. Consequently, strategies were developed to incorporate these unsure samples during network training.

The network utilized for this task is a WaveNet-inspired approach [Oord], which takes the centerline points starting from the ostium with a fixed length as input. To enhance the robustness of the model, techniques such as TTA and model ensembling are employed. Several approaches for handling unsure samples during network training are evaluated, including exclusion, random class assignment per ensemble and per individual model, and the assignment of a soft label. As we have the intuition that the non-confidently labeled cases form a distinct distribution within the probability

space over a test population, we assess whether abstaining from predictions for these samples would lead to a more robust performance. To accomplish this, we propose a quantile-based abstention rule that considers the a priori probability of SC RCA as an anchor and to preserve class balance after abstention.

Our results demonstrate an overall strong performance for the task at hand, achieving an area under the receiver operating characteristic curve (AUC) of 0.938 on the confidently labeled cases. We found that including the unsure samples in the training process, regardless of the method used, improved performance, with a slight advantage observed for the soft label assignment approach. Regarding our abstention rule, we observed better performance when considering the underlying label distribution instead of relying solely on the a priori probability provided in the literature, especially when excluding only a limited number of samples. Furthermore, we were able to confirm our assumption that the unsure cases form a distinct distribution, as the performance gap between the best and worst possible label assignment significantly decreased with an exclusion rate of 25 %.

In conclusion, our study successfully addressed the clinical need for automatic detection of SC RCA in screening scans. The developed WaveNet-inspired network, along with TTA and model ensembling, demonstrated strong performance. The inclusion of unsure samples during training, particularly through soft label assignment, proved beneficial. Additionally, our proposed quantile-based abstention rule, considering the label distribution, yielded improved results compared to relying solely on the a priori probability. These findings contribute to the advancement of automated screening for SC RCA, enhancing patient care and minimizing procedural complications.

7.2.2 Discussion

Within the project encompassing this thesis, the original motivation behind the research conducted in this paper is a much broader objective. As alluded to in Section 1.1.3 and also Section 3.1, there are numerous norm variants and anomalies of the coronary arteries and the heart in general. In clinical practice, physicians are knowledgeable about these variants as part of their training and these are highlighted when encountered in routine clinical scenarios. However, incorporating a focus on rare norm variants or anomalies into an artificial intelligence (AI) solution can be challenging. One of the primary reasons is the scarcity of data, which is even more pronounced for these exceptional cases. Consequently, evaluating the performance of CAD grading systems on these norm variants and anomalies is of significant interest. Some variants may have a substantial impact on the overall CAD analysis, such as duplicate RCA

or a missing left main (LM) segment. An automated system intended for clinical deployment must be able to handle these cases. However, given the data sparsity and the potential for earlier pipeline steps to fail in such cases, it becomes crucial to develop robust detection systems for these outliers. This can involve sanity checking the outcomes of the pre-processing steps. A final system should notify physicians that the confidence in the prediction for a given sample may be lower than usual.

Another important aspect derived from this research pertains to the handling of annotations in general. In our opinion, it is preferable to refrain from assigning a definitive label rather than assigning a non-confident one. We hope that such annotation practices find wider adoption in research, as they synergize with learning with abstention, where a similar line of thinking is applied to the neural network itself.

HANDLING LABEL UNCERTAINTY ON THE EXAMPLE OF AUTOMATIC DETECTION OF SHEPHERD'S CROOK RCA IN CORONARY CT ANGIOGRAPHY

Felix Denzinger[1,2], *Michael Wels*[2], *Oliver Taubmann*[2], *Florian Kordon*[1], *Fabian Wagner*[1], *Stephanie Mehltretter*[1], *Mehmet A. Gülsün*[2], *Max Schöbinger*[2], *Florian André*[3], *Sebastian Buß*[3], *Johannes Görich*[3], *Michael Sühling*[2], *Andreas Maier*[1], *Katharina Breininger*[4]

[1] Pattern Recognition Lab, FAU Erlangen-Nürnberg, Erlangen, Germany
[2] Siemens Healthcare GmbH, Computed Tomography, Forchheim, Germany
[3] Das Radiologische Zentrum, Sinsheim-Eberbach-Erbach-Walldorf-Heidelberg, Germany
[4] Department Artificial Intelligence in Biomedical Engineering, FAU Erlangen-Nürnberg

ABSTRACT

Coronary artery disease (CAD) is often treated minimally invasively with a catheter being inserted into the diseased coronary vessel. If a patient exhibits a Shepherd's Crook (SC) Right Coronary Artery (RCA) – an anatomical norm variant of the coronary vasculature – the complexity of this procedure is increased. Automated reporting of this variant from coronary CT angiography screening would ease prior risk assessment. We propose a 1D convolutional neural network which leverages a sequence of residual dilated convolutions to automatically determine this norm variant from a prior extracted vessel centerline. As the SC RCA is not clearly defined with respect to concrete measurements, labeling also includes qualitative aspects. Therefore, 4.23 % samples in our dataset of 519 RCA centerlines were labeled as unsure SC RCAs, with 5.97 % being labeled as sure SC RCAs. We explore measures to handle this label uncertainty, namely global/model-wise random assignment, exclusion, and soft label assignment. Furthermore, we evaluate how this uncertainty can be leveraged for the determination of a rejection class. With our best configuration, we reach an area under the receiver operating characteristic curve (AUC) of 0.938 on confident labels. Moreover, we observe an increase of up to 0.020 AUC when rejecting 10 % of the data and leveraging the labeling uncertainty information in the exclusion process.

Index Terms— Label Uncertainty, Shepherd's Crook RCA, Coronary CT Angiography

1. INTRODUCTION

Coronary artery disease (CAD) is an often deadly disease commonly linked to atherosclerotic plaque deposits narrowing the coronary vasculature [1]. These lesions are usually treated minimally invasively in the cath lab, where a catheter is inserted through the femoral artery. This catheter is then guided toward the location of the lesion. The vessel at the lesion is then widened and stabilized using a balloon and a stent. One anatomical norm variant of the coronary vasculature – the Shepherd's Crook right coronary ascending artery (SC RCA) – may complicate this procedure as the RCA branch takes a high and tortuous turn directly after the ostium (cf. Fig. 1 right) [2]. Furthermore, this variant is suspected to increase the risk of developing CAD [3]. Therefore, automated detection of SC RCA, e.g., from coronary CT angiography (CCTA) scans, is of interest. However, to the best of our knowledge, no prior work on this topic exists.

To develop a deep learning-based algorithmic solution, we build a data collection of 519 patients with labels indicating whether patients exhibit an SC RCA or not. However, as the sole definition of this norm variant is the high, tortuous turn, we did not only identify 31 cases we consider sure SC RCAs but also 22 border cases which we labeled as unsure, with an example displayed in Fig. 1. As these cases could not be labeled with high confidence by human readers, a machine learning approach should also rather not report a prediction for such samples instead of confidently predicting a label. Therefore, methods from uncertainty estimation or abstention learning are considered, where instead of just learning to distinguish presence from absence, a rejection class is additionally determined from the output of a machine learning model.

In summary, we formulate the following research questions:

1. Can we automatically determine whether a patient has a SC RCA using a data-driven algorithm?
2. How should samples be handled for which the annotator is not confident?
3. Can we leverage the labeling uncertainty to enhance or at least better evaluate learning with abstention?

We tackle them with the following contributions:

1. Development of a deep learning approach which analyzes the centerline course of the RCA using a WaveNet-like 1D convolutional neural network.

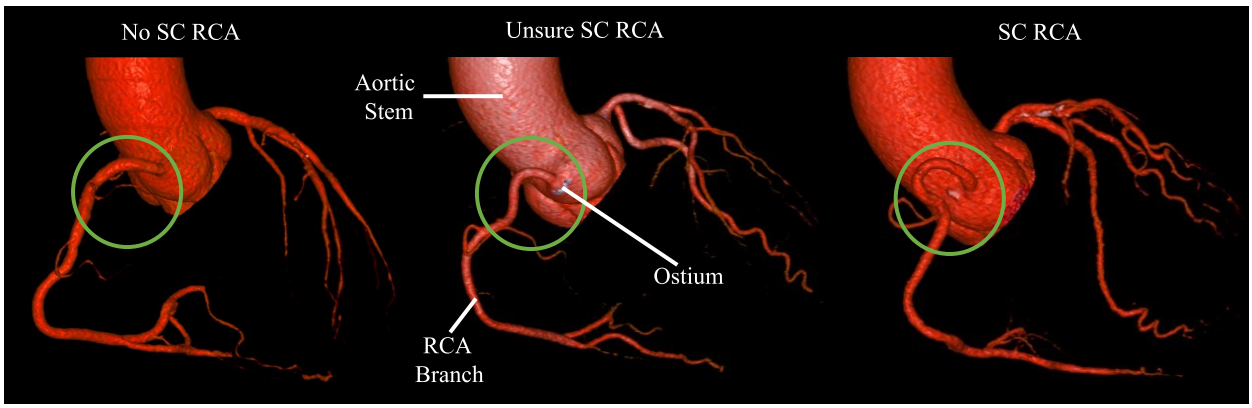

Fig. 1. Volume rendering of the aortic stem and coronary arteries of three different patients: left) patient without a Shepherd's Crook (SC) RCA, center) patient labeled as having an unsure SC RCA as the RCA does take a tortuous high turn but to a lesser extent, and right) patient with a SC RCA defined by a high, tortuous turn after the origin of the RCA segment.

2. Analysis of four different ways to handle the cases labeled as unsure: exclusion during training, randomly assigning a class either globally or for each model in an ensemble, or assigning a soft label.
3. Proposal of a non-invasive percentile-based rejection scheme and examination of whether information about the frequency of uncertain samples can improve it.

2. MATERIAL AND METHODS

2.1. Data

Within this study, a data collection of 519 CCTA scans is used. Of these, 31 (5.97 %) are labeled as positive SC RCA cases, and 22 (4.23 %) as unsure. Labeling was performed by a doctoral researcher with four years of experience in the field of CAD assessment from CCTA scans. Centerlines of these scans were extracted using the well-established and robust algorithm of Zheng et al. [4]. From these, the first 64 mm (256 points, spacing of 0.25 mm) of the RCA were extracted by combining the proximal and middle RCA segments as provided by the labeling algorithm of Denzinger et al. [5]. We consider the coordinates of the centerline as features that are normalized by subtracting the coordinates of the first centerline point (ostium) from all points and then dividing by 64 mm, as this is the maximal possible length of an input centerline.

2.2. Deep Learning-based Shepherd's Crook Detection

As the local and global curvature and the overall course of the centerline are key features to be determined by a classifier, we propose to use a WaveNet-inspired [6] deep learning architecture as depicted in Fig. 2. It leverages 1D convolutions of differing dilation grades to model short and long-range depen-

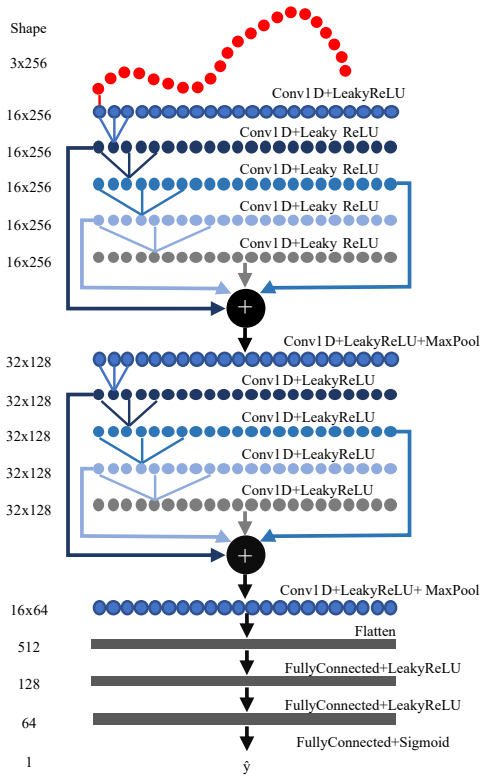

Fig. 2. Overview of the WaveNet-inspired 1D convolutional neural network for the classification of SC RCA. The input to the network are the 3D centerline coordinates, which are processed by a set of 1D convolutional layers with increasing dilation grade. The features created from different perceptive fields are summed up and fed into a second WaveNet-like block. The final feature representation is then processed by a multi-layer perceptron to predict the presence of SC RCA.

dencies, which are combined and weighted by a multi-layer perceptron [7].

To prevent overfitting, we randomly rotate our training data with the ostium as the rotation center in a range of up to 45 ° in all directions.Furthermore, we use a binary cross-entropy loss, an Adam optimizer with the default learning rate of 0.001, and default batch size of 32. At test time, data is augmented by rotations of $[-15\,^\circ, 0\,^\circ, 15\,^\circ]$, with the final prediction being the mean across all rotations. We use a fixed amount of 100 epochs to omit the need for a validation set due to the small number of SA RCA in the dataset. To improve the robustness of the prediction under this setting, we combine five training runs to form one final model by averaging the predictions of the five sub-models. Due to the limited amount of data, the performance statistics differed for repeated experiments. To obtain reproducible results, we performed a 5-fold cross-validation and repeated it 25 times. Data was split in a stratified manner regarding both positive and unsure samples.

2.3. Label Uncertainty Handling

In this work, we evaluate four different strategies to handle samples with unsure labels: randomly assigning them to one class globally ("Fixed") or for each training run ("Varied"), not including them in the training phase ("Exclusion"), or assigning a soft label of 0.5 ("0.5").

The global random assignment of all samples mimics the usual handling of unsure cases, which are, in practice, not labeled as such, but some class assignment is enforced. With the random assignment for each individual training run of the ensemble and then combining the prediction over 5 of these runs, the output probabilities for the unsure cases should lie in between the distributions of the sure cases.

Not including the unsure samples is also a valid strategy but decreases the amount of data seen by the network.

Assigning a soft label of 0.5 to the unsure samples encourages the probabilities of the network to form a separate distribution between the negative and positive classes.

2.4. Percentile-based Abstention

As there are samples marked as unsure, evaluating whether this labeling uncertainty can be confirmed by model uncertainty is an obvious choice. Therefore, we perform learning with abstention, i.e., determining which samples should be rejected. Because of the class imbalance, defining an abstention rule around the probability value of 0.5 (anchor) is not applicable. Instead, we propose a percentile-based approach: a frequently reported value for the prevalence $p(y_1)$ of the SC RCA is 5 %[8]. Therefore, for an ideal classifier the highest 5 % of the test-set predictions would belong to the positive class. To account for this, we run inference on the entire test set and select the prediction value at the 95th percentile (1-$p(y_1)$) in the probability histogram as the anchor of our abstention interval.

From our defined anchor, we define our exclusion interval as:

$$\begin{aligned} p_{\text{min}} &= 1 - p(y_1) - e * p(y_0) \\ p_{\text{max}} &= 1 - p(y_1) + e * p(y_1) \end{aligned} \tag{1}$$

with p_{min} and p_{max} also referring to percentile values in the histogram and $e \in [0; 1]$ denoting the exclusion rate, which can be varied to specify the amount of coverage, i.e., the amount of data kept after abstention. With this interval, we keep the class balance also after abstention, as samples are excluded in relation to the prior probability of both classes. Note that the probability values corresponding to the percentiles can be transferred to new single samples as well.

Additionally, we examine whether a better abstention interval can be achieved by additionally leveraging the information of the frequency of the unsure samples observed in the training data. To this end, we replace $p(y_1)$ with $p(y_1) + p(y_{0.5})$ and $p(y_0)$ with $p(y_0) - (p(y_1) + p(y_{0.5}))$ in the interval defined above. We call this configuration "$p(y_{0.5})$".

2.5. Evaluation

Since there are no ground truth labels for the unsure samples, we propose using the following three performance measures: we calculate the AUC for all possible permutations of class assignment for the unsure cases to get the best and worst possible AUC value. Additionally, we report the performance solely on the sure samples. As discussed in Section 2.2, 5-fold cross validation with 25 repetitions was used to obtain robust results given the small number of overall samples.

3. RESULTS

Our results for the different evaluated configurations are displayed in Table 1. From a high-level perspective, there is a relatively large gap between the best and worst possible AUC, indicating how much of an impact the relatively small number of unsure cases can have during test time. Generally, we reach excellent performance on the data set consisting of high confidence labels with an AUC of up to 0.940 at 100 % coverage.

The choice of how to handle unsure cases during training had a small effect. There is a clear trend that the exclusion of the borderline cases leads to a worse performance. Having a random assignment for each single training run or globally performed comparably, with a soft label of 0.5 performing best.

Regarding abstention, one can recognize that the metrics stayed mostly the same when only excluding 5 % of the data and increased slightly at 10 %. Paired with the observation that the distance between the best and worst possible AUC is not decreasing, it becomes apparent that only a limited amount of unsure samples lies in this initial exclusion interval.

Config	AUC	100 %	95 %	90 %	75 %
Exclusion	Best	0.942	0.942	0.946	0.963
Fixed	Best	0.944*	0.946	0.951	**0.971**
Varied	Best	0.944	0.945	0.950	0.969
0.5	Best	**0.945**	0.947	0.954	**0.971**
0.5 $p(y_{0.5})$	Best	**0.945**	**0.951**	**0.958**	0.970
Exclusion	Worst	0.878	0.874	0.877	0.922
Fixed	Worst	0.885***	0.885	0.890	0.939
Varied	Worst	0.885	0.885	0.892	0.940
0.5	Worst	**0.887**	0.887	0.894	0.941
0.5 $p(y_{0.5})$	Worst	**0.887**	**0.899**	**0.914**	**0.950**
Exclusion	Sure	0.934	0.931	0.933	0.954
Fixed	Sure	0.937**	0.937	0.941	0.965
Varied	Sure	0.937	0.937	0.941	0.964
0.5	Sure	**0.938***	0.939	0.945	0.966
0.5 $p(y_{0.5})$	Sure	**0.938**	**0.945**	**0.953**	**0.967**

Table 1. Performance with respect to the AUC for different handling of unsure cases with differing amount of coverage. "Best" and "Worst" are determined by calculating the AUC for all possible label assignments of the unsure samples. The "Sure" AUC is calculated only using the samples labeled with high confidence. Note that 0.5 and 0.5 $p(y_{0.5})$ refer to the same training configuration with different abstention parameterization. Significance was determined using paired-sample t-test. For 100 % coverage, the significance levels are displayed as * in relation to the next worst configuration with respect to the AUC with the following p-value thresholds: $^* := p < .05$, $^{**} := p < 0.01$ and $^{***} := p < 0.002$ $p^* := 0.002$ according to Bonferroni correction.

However, when including the information about the frequency of both true positive and unsure cases ($p(y_{0.5})$) in the determination of the abstention interval, we notice an improvement of up to 0.021 for the worst possible AUC at a coverage of 90 %. Also, the distance between the best and worst possible AUC decreases to a great extent, especially at a coverage of 75 %. This indicates that a majority of the unsure samples lies within this exclusion interval and therefore form a distinct distribution in the probability space.

4. DISCUSSION

First, we want to discuss the model design choice. We are unaware of any work performing classification regarding the course of the centerline. However, there are related works on registration or segment labeling of centerlines which either utilize 1D convolutions without dilation [9] or recurrent neural networks [10, 11]. We tested these approaches in initial experiments, but the model proposed in this manuscript yielded better results. Therefore, this WaveNet-like feature extraction might also be applicable to other approaches in this area.

Architectures like PointNet [12] or ones based on graph deep learning [13] are more complex alternatives for the task at hand. These approaches might face additional challenges due to overfitting on a global structure or struggle to learn the internal connectivity of the centerline but this could be explored in future work.

Regarding our second research question of how to handle unsure samples, two related research fields come into mind: combating label noise and how to merge multiple annotators. Methods to combat label noise might be able to improve results presented in this paper. However, these algorithms do not answer the question on how to handle the unsure cases in the first place. With multiple annotators, strategies like taking the majority vote as confident label or ensembles, for which every sub-model is trained on a different annotator exist. These concepts are very similar to the strategies we evaluated here. A similar concept for introducing soft labels from multiple annotators was proposed very recently [14]. However, there are no works linking this concept to abstention yet.

Regarding abstention, we propose a strategy that can be directly applied to a trained model. There are other methods for uncertainty estimation and abstention which either try to estimate an underlying gamma distribution in the probability space for each sample [15] or include a dedicated abstention class [16]. However, these are usually more invasive in that it needs more adaptations and a trained model cannot be taken as is.In contrast, the abstention strategy we propose is simple and non-invasive. The most similar approach we found directly predicts uncertainty as a side-task and performs percentile-based abstention based on this uncertainty output [17].

5. CONCLUSION

We tackled three research questions in our work: can we automatically detect SC RCA, how should unsure samples be handled and does the information on how many of these samples exist help to form a rejection class?

An affirmative answer to the first question is given by the strong performance of 0.938 AUC on the confidently labeled data. For the second research question we evaluated a set of strategies and found small differences in the performance yield. Overall, soft label assignment performed best. We hope that our work inspires others to perform similar analyses, especially as the concept of uncertain labels for single annotators is currently not widely applied in the medical domain. Finally, we proposed a percentile-based abstention strategy. Here, we showed that adding information regarding the frequency of unsure cases improved the performance by a large margin for different levels of coverage.

6. COMPLIANCE WITH ETHICAL STANDARDS

The CT examinations were clinically indicated by the referring physicians and conducted in accordance with current clinical standards, guidelines, and recommendations. The study was performed in accordance with the Declaration of Helsinki and was approved by the local ethics committee (S-226/2016 and S-758/2018, Ethikkommission der Medizinischen Fakultät Heidelberg, Germany). Subjects included as of January 2019 gave informed consent in the scientific data analyses. For the retrospective analyses of the datasets acquired before January 2019, a waiver of consent was granted by the aforementioned ethics committee.

7. ACKNOWLEDGMENTS

This work was partially funded by Siemens Healtcare GmbH, Erlangen, Germany.

K.B. gratefully acknowledges the support of d.hip campus - bavarian aim in form of a faculty endowment.

8. REFERENCES

[1] Valentin Fuster, Lina Badimon, Juan J Badimon, and James H Chesebro, "The pathogenesis of coronary artery disease and the acute coronary syndromes," *New England Journal of Medicine*, vol. 326, no. 5, pp. 310–318, 1992.

[2] Jabi E Shriki et al., "Identifying, characterizing, and classifying congenital anomalies of the coronary arteries," *Radiographics*, vol. 32, no. 2, pp. 453–468, 2012.

[3] Muzaffer Saglam et al., "Shepherd's crook right coronary artery: a multidetector computed tomography coronary angiography study," *Kardiologia Polska (Polish Heart Journal)*, vol. 73, no. 4, pp. 261–273, 2015.

[4] Yefeng Zheng, Huseyin Tek, and Gareth Funka-Lea, "Robust and accurate coronary artery centerline extraction in CTA by combining model-driven and data-driven approaches," in *International Conference on MICCAI*. Springer, 2013, pp. 74–81.

[5] Felix Denzinger et al., "CAD-RADS Scoring using Deep Learning and Task-Specific Centerline Labeling," in *MIDL*, 2021.

[6] Aäron van den Oord et al., "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, pp. 125–125.

[7] Andreas Maier, Christopher Syben, Tobias Lasser, and Christian Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 86–101, 2019.

[8] David E Gossman, E Murat Tuzcu, Conrad Simpfendorfer, and Gerald J Beck, "Percutaneous transluminal angioplasty for shepherd's crook right coronary artery stenosis," *Catheterization and Cardiovascular Diagnosis*, vol. 15, no. 3, pp. 189–191, 1988.

[9] Wei Wu et al., "CAR-Net: A Deep Learning-Based Deformation Model for 3D/2D Coronary Artery Registration," *IEEE Transactions on Medical Imaging*, 2022.

[10] Dan Wu et al., "Automated anatomical labeling of coronary arteries via bidirectional tree LSTMs," *IJCARS*, vol. 14, no. 2, pp. 271–280, 2019.

[11] A Fischer et al., "Deep learning based automated coronary labeling for structured reporting of coronary CT angiography in accordance with SCCT guidelines," *Journal of Cardiovascular Computed Tomography*, vol. 14, no. 3, pp. S21–S22, 2020.

[12] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings of the IEEE Conference on VCPR*, 2017, pp. 652–660.

[13] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen, "An end-to-end deep learning architecture for graph classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[14] Katherine M Collins, Umang Bhatt, and Adrian Weller, "Eliciting and learning with soft labels from every annotator," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2022, vol. 10.

[15] Florin C Ghesu et al., "Quantifying and leveraging classification uncertainty for chest radiograph assessment," in *International Conference on MICCAI*. Springer, 2019, pp. 676–684.

[16] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof, "Combating label noise in deep learning using abstention," in *International Conference on Machine Learning*, 2019, pp. 6234–6243.

[17] Elizabeth A Barnes and Randal J Barnes, "Controlled abstention neural networks for identifying skillful predictions for regression problems," *Journal of Advances in Modeling Earth Systems*, vol. 13, no. 12, 2021.

PART III

Outlook and Summary

CHAPTER 8

Outlook

Having presented the diverse range of research conducted within this thesis and establishing the broader context of grading coronary artery disease (CAD) from coronary CT angiography (CCTA) scans, it is essential to discuss potential future research directions and open questions.

Overall, CAD scoring systems aim to provide insight into individual patient risk and guide clinical decisions with the ultimate goal of mitigating major adverse cardiac events. However, directly predicting these events or deriving risk factors poses significant challenges, as patients at high risk typically receive intensive care. This in turn, leads to a decreased risk for the event actually manifesting which is a challenge known as the prevention paradox. Consequently, relying on established scoring systems used in clinical practice remains necessary. However, these scoring systems are continuously assessed for their effectiveness, and a combination of approaches is typically employed. A recent development in this field is the recommendation of coronary artery disease-reporting data system (CAD-RADS) score 2.0 by Cury et al. [Cury 22]. The main difference from the previous CAD-RADS is the inclusion of computed tomography (CT)-derived fractional flow reserve (FFR) to guide decision-making for patients with confirmed significant stenosis. Additionally, CT scans can directly assess heart muscle perfusion [Gonz 15], providing further insights into the haemodynamic impact of single lesions and the overall risk of a patient. Consequently, computer-assisted diagnosis systems need to incorporate various tasks and appropriately weigh their importance, including stenosis grades, haemodynamic significance, heart muscle perfusion, and the presence of stents or other grafts. This provides numerous challenges for future research, e.g., automated analysis and the exploration of synergies between these different aspects. Furthermore, leveraging information across multiple modalities may offer a more comprehensive view of patient risk.

To achieve these goals, it is crucial to aggregate data with both high variance and high-quality annotations. One potential approach is to deepen our understanding of the anatomy and variability of the heart and coronary arteries, in conjunction with

downstream imaging systems. This knowledge could facilitate the generation of large quantities of synthetic data, including rare anomalies and norm variants, to enhance performance on these challenging cases [Meis 20]. Image synthesis algorithms have gained attention and are already being applied to CT volume synthesis [Khad 23]. However, to be applicable for the tasks at hand, anatomically plausible image data, in a holistic manner, i.e., for the whole disease state, must be simulated. Additionally, generating a large volume of data necessitates a substantial need for data annotations, which remain expensive.

Moreover, the field is poised for significant transformation with ongoing technical advancements in the field of computed tomography CT. The emergence of photon counting CT has paved the way for various exciting research avenues. This includes the potential to gain a deeper understanding of vulnerable plaques through the utilization of ultra-high-resolution scans and material decomposition techniques. Additionally, it offers the possibility of improving the assessment of stenosis degree and CT-derived FFR by mitigating the impact of calcium blooming artifacts.

Several challenges persist regarding the work conducted in this thesis, including the explainability of network decisions, enhancing robustness against technical variations, developing a reliable approach to handle anatomical outliers, and enabling learning with abstention. As discussed throughout previous chapters, there is definite value in creating an explainable system that allows readers to easily understand the network's decision-making process. Various approaches have been suggested, such as highlighting the highest activating subsegment or displaying the angle with the highest outputted probability. However, these approaches require careful qualitative clinical evaluation for validation. Furthermore, as demonstrated in Section 7.1.3, both the pre-processing algorithms and the proposed artificial neural network (ANN) are sensitive to technical variations. As previously discussed, disentangling the biological from the technical information may offer potential solutions to overcome this limitation. Additionally, addressing norm variants and coronary anomalies requires robust detection algorithms to exclude them from further processing or specific collection to enhance the robustness of an automated artificial intelligence (AI) system.

Given the overall strong performance our method achieves, we assume that our model is currently limited by the availability of lesion-wise annotations and inter-/intra-operator agreement. To address the latter limitation, conducting studies with data collections annotated by multiple readers, or as we propose, annotations that include reader confidence, is necessary. These studies would enable us to assess how well our system performs in comparison to a trained radiologist and also give an intuition on the maximum achievable performance.

CHAPTER 9

Summary

To allow a better overview of the overall content of this thesis, it will wrap up with a summary. As seen, coronary artery disease (CAD) is an impactful disease, and research to automatically detect it is of high interest. It is usually related to atherosclerotic plaque deposits that narrow the coronary artery lumen, obstructing blood flow. The degree of stenosis caused by these lesions is frequently assessed in the clinics as it is an established indicator of the impact on the patient's health. A modality to examine these lesions in a non-invasive manner is coronary CT angiography (CCTA). From CCTA scans, the coronary artery disease-reporting data system (CAD-RADS) score is determined, which is mainly influenced by the most severe stenosis for a patient. The coronary artery is subdivided into three branches including the right coronary artery (RCA), left artery descending (LAD), and circumflex artery (CX). For the RCA a norm variant called Shepherd's crook (SC) RCA exists, which may complicate minimally invasive procedure due to its tortuous high turn after the ostium. Coming from this medical background, we define the main research question of this thesis being: *Can we create a computer-assisted diagnosis system to automatically determine CAD from CCTA data?* To answer this question, we conducted research in three areas: in the first part, we perform characterization of single plaque lesions with respect to significant stenosis degree and revascularization decision using deep learning (DL). Next, we propose DL-based methods to automatically determine the CAD-RADS score. Finally, we evaluate differing aspects of the clinical applicability of such methods.

To foster a superficial understanding of the concept of CCTA acquisition and reconstruction, Chapter 2 provides a brief basic overview of these topics. By acquiring X-ray projection images from multiple angles, a 3D computed tomography (CT) volume can be reconstructed by leveraging the filtered back projection (FBP) algorithm. For CCTA acquisition specifically, there are some important concepts including the parameterization of the FBP algorithm and how neighboring imaging stacks acquired at different heart cycles are combined. Parameters to select include the reconstruction kernel and the number of iterations for the advanced modeled iterative reconstruction

(ADMIRE) algorithm, which is an iterative FBP variant.

Next, some basic concepts behind machine learning (ML) are elaborated on in Chapter 3. These are the foundation of the methods applied in this thesis. In a first step, the main terminologies including measurements, features, predictions and ground truth are introduced together with the pattern recognition pipeline. The latter describes the workflow on how a ML system is trained and then deployed during inference. Some key challenges of ML in medical image analysis include the scarcity of data and that annotations are not always without variance, i.e., there may be some inherent label uncertainty for hard-to-label cases. Furthermore, medical image analysis tasks usually suffer from class imbalance and there are often prevalence and domain shifts for different populations and sites. The chapter continues with a brief recap of decision trees – which form a simple rule-based classifier – and an extension of them named eXtreme Gradient Boosting (XGBoost). The latter leverages ensembling of multiple decision trees where new trees are added based on greedy search. Next, DL-based ML is explained where multiple input features are processed by a series of functions. Commonly, scaling functions and non-linear activation functions are alternated, which enables the approximation of any function. By propagating the gradient with respect to a suitable loss function backwards through this chain of functions – leveraging the chain rule on the way – the gradient with respect to the weight parameters of the scaling functions can be calculated. These gradients can then be used to update the weights. This same principle can be applied on images using convolutional neural networks (CNNs), and on time series data using recurrent neural networks (RNNs).

Due to a high quantity of novel literature in the field, it rendered necessary to provide an overview of the most recent related work in Chapter 4. First, intuition on how CAD is assessed in the clinical workflow is fostered. The usual pipeline leverages already established pre-processing including centerline extraction and labeling to help with the localization of lesions. Furthermore, multi planar reformation (MPR) and curved planar reformation (CPR) images which utilize the centerlines for efficient visualization of the area of interest are employed. To narrow down the problem and input a good data representation to any artificial neural network (ANN) approach, most related work leverage the MPR or CPR views as an input. Furthermore, to incorporate the prior knowledge on the cylindrical structure of vessels in the MPR format, polar transformation can be leveraged. There exists a variety of architectures employed. The most commonly used method design utilizes local feature extractors along the centerline and then analyzes the sequence information using RNNs.

With the theoretical foundation and related work defined, research work on coronary

plaque characterization is presented in Chapter 5. Here, the targets are significant stenosis grade and the revascularization decision on lesion-level. In the first presented paper, three different approaches are proposed. The first approach takes a segmentation of a coronary plaque as input, calculates several different features from the mask and CCTA volume and then predicts the targets at hand using the XGBoost classifier. A second approach extracts features from MPR cubes along the centerline transformed into a cylindrical coordinate system and analyzes the resulting sequence using an RNN. As a final approach, a combination of the first two approaches is proposed, where masks along the centerline are segmented and for each mask, shape features are extracted. The sequence of features is then analyzed using an RNN. As two of these approaches require a prior segmentation, which is an additional error-prone step, further research is conducted. Here, an efficient data representation is leveraged where only two orthogonal longitudinal slices along the MPR volume are taken as an input. Additionally, a lightweight CNN and test time augmentation (TTA) are employed. Overall, the choice of input representation and architectural design are motivated by the characteristics of the pathology at hand. We are able to find a well-working data representation and network design. Result-wise the best performing method of our first paper reaches an area under the receiver operating characteristic curve (AUC) of 0.96 and 0.88 for the tasks of significant stenosis and revascularization decision, respectively. However, this method had the need for prior segmentation as a major drawback. With the best segmentation-agnostic method, we are able to reach an AUC of 0.92 and 0.90 on the respective stenosis degree and revascularization decision classification tasks.

In the subsequent Chapter 6, learnings from the previous chapter are leveraged to enable patient-wise CAD-RADS grading. Here, a multi-step pipeline is proposed. First, the coronary centerlines are extracted and subsequently divided into subsegments according to the american heart association (AHA) guidelines with respective ML approaches. Then, MPR volumes for each subsegment are extracted and interpolated to one common length. Subsequently, a hierarchical network architecture is leveraged, which takes orthogonal longitudinal views cut from the MPR volumes as utilized for our prior research as input. For each segment, features are extracted using a CNN. These features are then combined using a global max pooling operation to have a global feature representation, which is next used to predict the CAD-RADS grade with a multi-layer perceptron (MLP). Apart from this, the segment-wise features are used to predict the stenosis grade as an auxiliary target. An additional auxiliary target is the Agatston score (AS) at patient-level. To get rid of the interpolation step necessary due to the segments of varying sizes and to have a more

robust centerline labeling altogether, we published a further paper on this topic. A simple rule-based centerline labeling approach is proposed, along with some minor improvements including a label encoding suited for ordinal classification and TTAs. This centerline labeling approach subdivides the coronary artery tree into segments of the same size. These may not be anatomically correct, but it enables comparison across patients and allows us to get rid of the interpolation step. With the final approach, we manage to rule-out patients to have CAD with an AUC of 0.942 and hold-out patients from further invasive assessment with an AUC of 0.950. These demonstrated strong results may allow for clinical application in a screening or second reader scenario.

However, there are some questions regarding the clinical applicability we evaluate in Chapter 7. As our proposed method is embedded into a larger pipeline, it is essential to evaluate its robustness with respect to deviation in the earliest pipeline steps. As one of the first steps to allow for CCTA interpretation is the CCTA reconstruction, we evaluate the influence of a set of clinically often varied image formation parameters on the final network prediction. Namely, we alter the ADMIRE strength, the choice of reconstruction kernel, and the stacking strategy. In order to disentangle the influence of these variations on the pre-processing steps from our ANN prediction, we also evaluate the performance changes with the pre-processing results propagated from a default configuration to all others. We observe that the overall performance of our method does not show a significant drop for all but the true stack configuration. This can be qualitatively explained, as the resulting stacking artifacts may lead to vessels appearing narrowed. Moreover, we observed changed predictions on a per-patient level. As one expects the prediction of a CAD system to mainly rely on biological information, we conclude that future work needs to disentangle the biological from the technical variation. Another aspect of the clinical applicability is how rarer norm variants or anomalies are handled. As a proof of concept, we propose a method to predict the SC RCA from centerline data. Since we find that assigning a confident label is not possible for all of the samples, we experiment with a border case class and tried to exclude samples from it using a strategy to abstain from assigning a class to some of the samples.

Moving forward from this, we outline various future research directions in Chapter 8. In a broader scope, they include the combination of different clinical scores and the synthesis of large quantities of data with generative models. For the methods presented in this work in particular, we identify the disentanglement of biological and technical variation, detection of anatomical outliers, explainability of the predictions, and a comparison with the maximum human performance as open topics.

List of Acronyms

ACC

accuracy 35

ADMIRE

advanced modeled iterative reconstruction 19, 20, 94, 119, 122

AHA

american heart association 6, 9, 38, 67, 68, 121

AI

artificial intelligence 8, 108, 118

ANN

artificial neural network 9, 25, 28–30, 33, 34, 79, 94, 118, 120, 122

AS

Agatston score 5, 67, 68, 121

AUC

area under the receiver operating characteristic curve III, V, 36, 47, 58, 69, 80, 108, 121, 122

CAD

coronary artery disease III, 3, 4, 6, 8, 15, 37, 46, 66, 93, 94, 108, 117, 119, 120, 122

CAD-RADS

coronary artery disease-reporting data system III, V, 6, 8–10, 41, 67–69, 91–94, 117, 119, 121

CCS

coronary calcium scoring 5

CCTA

coronary CT angiography III, 5, 8, 15, 17, 20, 21, 24, 25, 37–39, 91, 94, 117, 119, 121, 122

CNN

convolutional neural network 9, 40, 41, 46, 47, 120, 121

CPR

curved planar reformation 38, 39, 41, 91, 120, 133

CT

computed tomography 5, 11, 17, 18, 24, 25, 94, 117–119

CV

computer vision 25

CX

circumflex artery 7, 91, 119

DL

deep learning III, V, 8, 15, 23, 25, 26, 28, 46, 57, 67, 80, 93, 119, 120

ECG

electrocardiogram 21

FBP

filtered back projection 18–20, 24, 119, 120, 133

FFR

fractional flow reserve 4, 6, 117, 118

FN

false negative 35

FP

false positive 35, 36

GMM

Gaussian mixture model 40

GRU

gated recurrent unit 34, 35, 40, 133

HU

Hounsfield unit 5, 19, 24

ILSVRC

ImageNet Large Scale Visual Recognition Challenge 25

IVUS

intravascular ultrasound 4, 5

KCTA

Koronar CT Angiographie V

KHK

koronare Herzkrankheit V, VI

LAD

left artery descending 7, 91, 119

LM

left main 7, 109

LSTM

long short term memory 34, 35, 40

MCC

Matthew's correlation coefficient 36, 69, 80

ML

machine learning III, V, 8, 10, 15, 23–26, 36, 120, 121

MLP

multi-layer perceptron 34, 121

MPR

multi planar reformation 38–40, 47, 57, 69, 91, 120, 121, 133

MSE

mean squared error 29

OCTA

optical coherence tomography angiography 4, 5

R-CNN

region-based convolutional neural networks 40, 41

RCA

right coronary artery 7, 10, 107, 108, 119, 122

ReLU

rectified linear unit 31, 32

RI

ramus intermedius 7, 91

RNN

recurrent neural network 9, 34, 40, 46, 47, 120, 121

SC

Shepherd's crook 7, 10, 107, 108, 119, 122

SD

stenosis degree 6

SVM

support vector machine 40

TN

true negative 35

TP

true positive 35, 36

TTA

test time augmentation 9, 57, 58, 65, 79, 80, 91, 95, 107, 108, 121, 122

XGBoost

eXtreme Gradient Boosting 26, 27, 46, 120, 121, 131

List of Symbols

General Notation

Symbol	Description
a	Coordinate
a	Scalar
$f()$	Function
$\mathbf{a}$	Vector
$\mathbf{A}$	Matrix
A	Set
max	Maximum value
cos	Cosine function
da	Differential with respect to a
∂	Partial derivative
log	Logarithm with base 10
μ	Mean value
σ	Standard deviation

Introduction

Symbol	Description
l	Lesion index
s_l	Lesion-wise calcification score
v_l	Lesion volume
w_d	Density weight

$\mathbf{C_L}$	Intensities of the lesion segmentation

Coronary CT Angiography

Symbol	Description
I	Beam intensity
I_0	Initial beam intensity
ζ	Attenuation coefficient
θ	Rotation angle
$p_\theta(\mathrm{s})$	Line projection
$F(\mathrm{u}, \mathrm{v})$	Function in Fourier space
$f(\mathrm{x}, \mathrm{y})$	Function in 2D space
ξ	Frequency variable in Fourier space
$P(\xi, \theta)$	1D Fourier transform of $p_\theta(s)$
$h(\mathrm{s})$	Inverse of ξ
Δ	Laplace operator
$H(\omega)$	Fourier transform of $h(\mathrm{s})$
$\omega_{\max}$	Reconstruction kernel positive width
$\omega_{\min}$	Reconstruction kernel negative width

Machine Learning

Symbol	Description
$\boldsymbol{\omega}$	Information vector
$\boldsymbol{\Omega}$	Real world
m	Measurement
$\mathbf{m}$	Measurement vector
$\mathbf{m}'$	Pre-processed measurement vector

x	Feature
$\mathbf{x}$	Feature vector
$\hat{y}$	Prediction
$\hat{\mathbf{y}}$	Prediction vector
y	Ground truth
$\mathbf{y}$	Ground truth vector
$\boldsymbol{\alpha}$	Parameters of a classifier
$\mathcal{L}$	Loss function
$\mathcal{L}_{\text{XGB}}$	Loss function for the XGBoost classifier
t	Training iteration
i	Sample index
$\gamma()$	Regularization function
$g()$	Decision tree classifier function
$\mathcal{L}_{\text{diff}}$	Differentiable loss function
τ	Number of decision tree leaves
w_τ	Decision tree leaf weight
w_{reg}	Regularization weight
$\mathbf{w}$	Weight vector
$\mathcal{L}_{\text{BCE}}$	Binary cross entropy loss function
$\mathcal{L}_{\text{MSE}}$	Mean squared error loss function
ν	Learning rate
∇	Nabla operator
$\mathbf{p}$	Momentum
β_p	Momentum weight

$\mathbf{v}$	Velocity
β_v	Velocity weight
$\mathcal{A}$	Activation function
$\odot$	Element-wise multiplication
$\mathbf{W}$	Weight matrix
$\mathbf{I}$	Input image
$\mathbf{H}$	Convolution kernel
$*$	Convolution operation
d	Dilation grade
$\mathbb{N}$	Natural numbers
z	Update gate
r	Reset gate
h'	Proposed hidden state
h	Hidden state

List of Figures

Copyright Notice

Bibliography

[Agat 90] A. S. Agatston, W. R. Janowitz, F. J. Hildner, N. R. Zusmer, M. Viamonte, and R. Detrano. "Quantification of coronary artery calcium using ultrafast computed tomography". *Journal of the American College of Cardiology*, Vol. 15, No. 4, pp. 827–832, 1990.

[Aust 75] W. G. Austen, J. E. Edwards, R. L. Frye, G. Gensini, V. L. Gott, L. S. Griffith, D. C. McGoon, M. Murphy, and B. B. Roe. "A reporting system on patients evaluated for coronary artery disease. Report of the Ad Hoc Committee for Grading of Coronary Artery Disease, Council on Cardiovascular Surgery, American Heart Association". *Circulation*, Vol. 51, No. 4, pp. 5–40, 1975.

[Bohu 07] C. M. Bohun, J. E. Potts, B. M. Casey, and G. G. Sandor. "A population-based study of cardiac malformations and outcomes associated with dextrocardia". *The American Journal of Cardiology*, Vol. 100, No. 2, pp. 305–309, 2007.

[Cand 20] S. Candemir, R. D. White, M. Demirer, V. Gupta, M. T. Bigelow, L. M. Prevedello, and B. S. Erdal. "Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network". *Computerized Medical Imaging and Graphics*, Vol. 83, p. 101721, 2020.

[Chau 16] A. S. Chauhan and K. Mukherjee. "Economic burden of coronary heart disease in North India". *International Journal of Noncommunicable Diseases*, Vol. 1, No. 1, pp. 18–25, 2016.

[Chen 16] T. Chen and C. Guestrin. "XGBoost: a scalable tree boosting system". In: *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[Chen 22] J.-T. Chen, Y.-C. Huang, H. Roth, D. Yang, C.-K. Lee, W.-J. Lee, T.-D. Wang, C.-Y. Chou, and W. Wang. "Detection and Classification of Coronary Artery Plaques in Coronary Computed Tomography Angiography Using 3D CNN". In: *12th International Workshop, STACOM 2021*, pp. 208–218, Springer, 2022.

[Chol 17] F. Chollet. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

[Cury 16] R. C. Cury, S. Abbara, S. Achenbach, A. Agatston, D. S. Berman, M. J. Budoff, K. E. Dill, J. E. Jacobs, C. D. Maroules, G. D. Rubin, F. J. Rybicki, U. J. Schoepf, L. J. Shaw, A. E. Stillman, C. S. White, P. K. Woodard, and J. A. Leipsic. "Coronary artery disease-reporting and data system (CAD-RADS) an expert consensus document of SCCT, ACR and NASCI: endorsed by the ACC". *JACC: Cardiovascular Imaging*, Vol. 9, No. 9, pp. 1099–1113, 2016.

[Cury 22] R. C. Cury, J. Leipsic, S. Abbara, S. Achenbach, D. Berman, M. Bittencourt, M. Budoff, K. Chinnaiyan, A. D. Choi, B. Ghoshhajra, J. Jacobs, L. Koweek, J. Lesser, C. Maroules, G. D. Rubin, F. J. Rybicki, L. J. Shaw, M. C. Williams, E. Williamson, C. S. White, T. C. Villines, and R. Blankstein. "CAD-RADS™ 2.0–2022 Coronary Artery Disease-Reporting and Data System: An Expert Consensus Document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Cardiology (ACC), the American College of Radiology (ACR), and the North America Society of Cardiovascular Imaging (NASCI)". *Cardiovascular Imaging*, Vol. 15, No. 11, pp. 1974–2001, 2022.

[Cybe 92] G. Cybenko. "Approximation by superpositions of a sigmoidal function.". *Math. Control. Signals Syst.*, Vol. 5, No. 4, p. 455, 1992.

[Denz 19] F. Denzinger, M. Wels, N. Ravikumar, K. Breininger, A. Reidelshöfer, J. Eckert, M. Sühling, A. Schmermund, and A. Maier. "Coronary artery plaque characterization from CCTA scans using deep learning and radiomics". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 593–601, Springer, 2019.

[Denz 20a] F. Denzinger, M. Wels, K. Breininger, M. A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Sühling, and A. Maier. "Automatic CAD-RADS scoring using deep learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 45–54, Springer, 2020.

[Denz 20b] F. Denzinger, M. Wels, K. Breininger, A. Reidelshöfer, J. Eckert, M. Sühling, A. Schmermund, and A. Maier. "Deep learning algorithms for coronary artery plaque characterisation from CCTA scans". In: *Bildverarbeitung für die Medizin 2020*, pp. 193–198, Springer, 2020.

[Denz 21a] F. Denzinger, M. Wels, C. Hopfgartner, J. Lu, M. Schöbinger, A. Maier, and M. Sühling. "Coronary Plaque Analysis for CT Angiography Clinical Research". In: *Bildverarbeitung für die Medizin 2021*, pp. 223–228, Springer, 2021.

[Denz 21b] F. Denzinger, M. Wels, O. Taubmann, M. A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Suehling, and A. Maier. "CAD-RADS Scoring using Deep Learning and Task-Specific Centerline Labeling". In: *Medical Imaging with Deep Learning*, 2021.

[Denz 23a] F. Denzinger, M. Wels, K. Breininger, O. Taubmann, A. Mühlberg, T. Allmendinger, M. A. Gülsün, M. Schöbinger, F. André, S. J. Buss, J. Görich, M. Sühling, and A. Maier. "How scan parameter choice affects deep learning-based coronary artery disease assessment from computed tomography". *Scientific Reports*, Vol. 13, No. 1, p. 2563, 2023.

[Denz 23b] F. Denzinger, M. Wels, O. Taubmann, F. Kordon, F. Wagner, S. Mehltretter, M. A. Gülsün, M. Schöbinger, F. André, S. Buß, J. Görich, M. Sühling, A. Maier, and K. Breininger. "Handling Label Uncertainty on the Example of Automatic Detection of Shepherd's Crook RCA in Coronary CT Angiography". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023.

[Doer 23] S. Doerrich, F. Kordon, F. Denzinger, J. S. El Barbari, M. Privalov, S. Y. Vetter, A. Maier, and H. Kunze. "Fast 3D YOLOv3 based standard plane regression of vertebral bodies in intra-operative CBCT volumes". *Journal of Medical Imaging*, Vol. 10, No. 3, p. 034503, 2023.

[Gerb 23] A. Gerbasi, A. Dagliati, G. Albi, M. Chiesa, D. Andreini, A. Baggiano, S. Mushtaq, G. Pontone, R. Bellazzi, and G. Colombo. "CAD-RADS scoring of coronary CT angiography with Multi-Axis Vision Transformer: a clinically-inspired deep learning pipeline". *arXiv preprint arXiv:2304.07277*, 2023.

[Gonz 15] J. A. Gonzalez, M. J. Lipinski, L. Flors, P. W. Shaw, C. M. Kramer, and M. Salerno. "Meta-analysis of diagnostic performance of coronary computed tomography angiography, computed tomography perfusion, and computed tomography-fractional flow reserve in functional myocardial ischemia assessment versus invasive fractional flow reserve". *The American Journal of Cardiology*, Vol. 116, No. 9, pp. 1469–1478, 2015.

[Good 16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[Grah 14] B. Graham. "Fractional max-pooling". *arXiv preprint arXiv:1412.6071*, 2014.

[Gros 09] S. Grosskopf, C. Biermann, K. Deng, and Y. Chen. "Accurate, fast, and robust vessel contour segmentation of CTA using an adaptive self-learning edge model". In: *Medical Imaging 2009: Image Processing*, pp. 1413–1420, SPIE, 2009.

[Guls 14] M. A. Gülsün, G. Funka-Lea, Y. Zheng, and M. Eckert. "CTA coronary labeling through efficient geodesics between trees using anatomy priors". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 521–528, Springer, 2014.

[Gupt 20] V. Gupta, M. Demirer, M. Bigelow, K. J. Little, S. Candemir, L. M. Prevedello, R. D. White, T. P. O'Donnell, M. Wels, and B. S. Erdal. "Performance of a deep neural network algorithm based on a small medical image dataset: incremental impact of 3D-to-2D reformation combined with novel data augmentation, photometric conversion, or transfer learning". *Journal of Digital Imaging*, Vol. 33, pp. 431–438, 2020.

[Hamp 22] N. Hampe, S. G. van Velzen, R. N. Planken, J. P. Henriques, C. Collet, J.-P. Aben, M. Voskuil, T. Leiner, and I. Išgum. "Deep learning-based detection of functionally significant stenosis in coronary CT angiography". *Frontiers in Cardiovascular Medicine*, Vol. 9, 2022.

[Hast 09] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2, Springer, 2009.

[He 22] D. He, Y. Chen, and T. Liu. "Automated plaque and stenosis characterization using attention-based deep learning network on coronary CT angiography images". In: *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering*, pp. 1–6, IEEE, 2022.

[Ioff 15] S. Ioffe and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *Proceedings of the International Conference on Machine Learning*, pp. 448–456, PMLR, 2015.

[Jin 22] X. Jin, Y. Li, F. Yan, Y. Liu, X. Zhang, T. Li, L. Yang, and H. Chen. "Automatic coronary plaque detection, classification, and stenosis grading using deep learning and radiomics on computed tomography angiography images: a multi-center multi-vendor study". *European Radiology*, Vol. 32, No. 8, pp. 5276–5286, 2022.

[Khad 23] F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baeßler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J. N. Kather, and D. Truhn. "Denoising diffusion probabilistic models for 3D medical image generation". *Scientific Reports*, Vol. 13, No. 1, p. 7303, 2023.

[Kiri 13] H. Kirişli *et al.* "Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography". *Medical Image Analysis*, Vol. 17, No. 8, pp. 859–876, 2013.

[Kunz 22] S. Kunzmann, C. Marzahl, F. Denzinger, C. Bertram, R. Klopfleisch, K. Breininger, V. Christlein, and A. Maier. "First Steps on Gamification of Lung Fluid Cells Annotations in the Flower Domain". In: *Bildverarbeitung für die Medizin 2022*, pp. 223–228, Springer, 2022.

[Lamb 12] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, and A. Dekker. "Radiomics: extracting more information from medical images using advanced feature analysis". *European journal of cancer*, Vol. 48, No. 4, pp. 441–446, 2012.

[Lang 23] S. Langer, O. Taubmann, F. Denzinger, A. Maier, and A. Mühlberg. "DeepTechnome: Mitigating Unknown Bias in Deep Learning Based Assessment of CT Images". In: *Bildverarbeitung für die Medizin 2023: Proceedings, German Workshop on Medical Image Computing, Braunschweig*, pp. 177–182, Springer, 2023.

[LeCu 89a] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. "Handwritten digit recognition with a back-propagation network". *Advances in Neural Information Processing Systems*, Vol. 2, 1989.

[LeCu 89b] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation applied to handwritten zip code recognition". *Neural Computation*, Vol. 1, No. 4, pp. 541–551, 1989.

[Lin 22] A. Lin, N. Manral, P. McElhinney, A. Killekar, H. Matsumoto, J. Kwiecinski, K. Pieszko, A. Razipour, K. Grodecki, C. Park, *et al.* "Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study". *The Lancet Digital Health*, Vol. 4, No. 4, pp. e256–e265, 2022.

[Liu 23] C. Liu, L. Folle, F. Denzinger, and A. Maier. "Whole-body Multi-organ Segmentation using Distance Attention". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023.

[Llin 88] R. R. Llinás. "The intrinsic electrophysiological properties of mammalian neurons: insights into central nervous system function". *Science*, Vol. 242, No. 4886, pp. 1654–1664, 1988.

[Luga 14] F. Lugauer, Y. Zheng, J. Hornegger, and B. M. Kelm. "Precise lumen segmentation in coronary computed tomography angiography". In: *International MICCAI Workshop on Medical Computer Vision*, pp. 137–147, Springer, 2014.

[Ma 21] X. Ma, G. Luo, W. Wang, and K. Wang. "Transformer network for significant stenosis detection in CCTA of coronary arteries". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 516–525, Springer, 2021.

[Maas 13] A. L. Maas, A. Y. Hannun, and A. Y. Ng. "Rectifier nonlinearities improve neural network acoustic models". In: *Proceedings of the International Conference on Machine Learning*, p. 3, Atlanta, Georgia, USA, 2013.

[Maie 18] A. Maier, S. Steidl, V. Christlein, and J. Hornegger. "Medical imaging systems: An introductory guide". 2018.

[Maie 19a] A. Maier, C. Syben, T. Lasser, and C. Riess. "A gentle introduction to deep learning in medical image processing". *Zeitschrift für Medizinische Physik*, Vol. 29, No. 2, pp. 86–101, 2019.

[Maie 19b] A. K. Maier, C. Syben, B. Stimpel, T. Würfl, M. Hoffmann, F. Schebesch, W. Fu, L. Mill, L. Kling, and S. Christiansen. "Learning with known operators reduces maximum error bounds". *Nature Machine Intelligence*, Vol. 1, No. 8, pp. 373–380, 2019.

[Mart 20] C. Martín Vicario, F. Kordon, F. Denzinger, M. Weiten, S. Thomas, L. Kausch, J. Franke, H. Keil, A. Maier, and H. Kunze. "Automatic plane adjustment of orthopedic intraoperative flat panel detector CT-volumes". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 486–495, Springer, 2020.

[Mart 22a] C. Martín Vicario, F. Kordon, F. Denzinger, J. S. El Barbari, M. Privalov, J. Franke, A. Maier, and H. Kunze. "Normalization techniques for CNN based analysis of surgical cone beam CT volumes". In: *Medical Imaging 2022: Image Processing*, pp. 648–652, SPIE, 2022.

[Mart 22b] C. Martín Vicario, F. Kordon, F. Denzinger, J. S. El Barbari, M. Privalov, J. Franke, S. Thomas, L. Kausch, A. Maier, and H. Kunze. "Automatic plane adjustment of orthopedic intraoperative flat panel detector CT-volumes". *Journal of Medical Imaging*, Vol. 9, No. 3, pp. 034001–034001, 2022.

[Maur 14] P. Maurovich-Horvat, M. Ferencik, S. Voros, B. Merkely, and U. Hoffmann. "Comprehensive plaque assessment by coronary CT angiography". *Nature Reviews Cardiology*, Vol. 11, No. 7, pp. 390–402, 2014.

[Meis 20] F. Meister, H. Houle, C. Nita, A. Puiu, L. M. Itu, and S. Rapaka. "Additional clinical applications". In: *Artificial Intelligence for Computational Modeling of the Heart*, pp. 183–210, Elsevier, 2020.

[Mose 23] P. T. Moser, R. Schernthaner, C. Loewe, A. Strassl, F. Denzinger, S. Faby, M. Wels, V. Nizhnikava, K. Uyanik-Uenal, A. Zuckermann, M.-E. Stelzmüller, and D. Beitzke. "Evaluation of perivascular fat attenuation with coronary CT angiography in cardiac transplantation patients: an imaging biomarker candidate for prediction of cardiac mortality and re-transplantation". *European Radiology*, pp. 1–9, 2023.

[Niem 13] H. Niemann. *Klassifikation von Mustern.* Springer-Verlag, 2013.

[Oord] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". In: *9th ISCA Speech Synthesis Workshop*, pp. 125–125.

[Pabl 19] A. Tejero-de Pablos, K. Huang, H. Yamane, Y. Kurose, Y. Mukuta, J. Iho, Y. Tokunaga, M. Horie, K. Nishizawa, Y. Hayashi, *et al.* "Texture-based classification of significant stenosis in CCTA multi-view images of coronary arteries". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 732–740, Springer, 2019.

[Pack 23] K. Packhäuser, S. Gündel, F. Thamm, F. Denzinger, and A. Maier. "Deep Learning-based Anonymization of Chest Radiographs: A Utility-preserving Measure for Patient Privacy". In: *Accepted at the International Conference on Medical Image Computing and Computer-Assisted Intervention*, p. , Springer, 2023.

[Paul 22] J.-F. Paul, A. Rohnean, H. Giroussens, T. Pressat-Laffouilhere, and T. Wong. "Evaluation of a deep learning model on coronary CT angiography for automatic stenosis detection". *Diagnostic and Interventional Imaging*, Vol. 103, No. 6, pp. 316–323, 2022.

[Pens 23] M. Penso, S. Moccia, E. G. Caiani, G. Caredda, M. L. Lampus, M. L. Carerj, M. Babbaro, M. Pepi, M. Chiesa, and G. Pontone. "A token-mixer architecture for CAD-RADS classification of coronary stenosis on multiplanar reconstruction CT images". *Computers in Biology and Medicine*, Vol. 153, p. 106484, 2023.

[Pfaf 23] L. Pfaff, F. Wagner, J. Hossbach, L. Pfaff, E. Preuhs, N. Maul, M. Thiess, F. Denzinger, M. Nickel, T. Wuerfl, and A. Maier. "Robust Multi-Contrast MRI Denoising Using Trainable Bilateral Filters Without Noise-Free Targets". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023.

[Pijl 96] N. H. Pijls, B. de Bruyne, K. Peels, P. H. van der Voort, H. J. Bonnier, J. Bartunek, and J. J. Koolen. "Measurement of fractional flow reserve to assess the functional severity of coronary-artery stenoses". *New England Journal of Medicine*, Vol. 334, No. 26, pp. 1703–1708, 1996.

[Quer 23] A. von Querfurth, F. Kordon, F. Denzinger, K. Breininger, and H. Kunze. "Lung, nodule and airway segmentation using partially annotated data". In: *Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling*, pp. 409–418, SPIE, 2023.

[Rami 18] J. C. Ramirez-Giraldo, K. L. Grant, and R. Raupach. "Admire: Advanced modeled iterative reconstruction". *Siemens Healthcare White Paper*, 2018.

[Rist 23] L. Rist, O. Taubmann, A. Mühlberg, F. Denzinger, F. Thamm, Nörenberg, J. Holch, S. Maurus, L. Gebauer, T. Huber, and A. Maier. "Spatial Lesion Graphs: Analyzing Liver Metastases with Geometric Deep Learning for Cancer Survival Regression". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023.

[Rohl 23] M. Rohleder, C. Pradel, F. Wagner, M. Thies, N. Maul, F. Denzinger, A. Maier, and B. Kreher. "Enabling Geometry Aware Learning Through

Differentiable Epipolar View Translation". In: *Accepted at the International Conference on Medical Image Computing and Computer-Assisted Intervention*, p. , Springer, 2023.

[Role 14] T. Roleder, J. C. Kovacic, Z. Ali, R. Sharma, E. Cristea, P. Moreno, S. K. Sharma, J. Narula, and A. S. Kini. "Combined NIRS and IVUS imaging detects vulnerable plaque using a single catheter system: a head-to-head comparison with OCT.". *Eurointervention: Journal of EuroPCR in Collaboration with the Working Group on Interventional Cardiology of the European Society of Cardiology*, Vol. 10, No. 3, pp. 303–311, 2014.

[Rose 58] F. Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain.". *Psychological review*, Vol. 65, No. 6, p. 386, 1958.

[Roth 20] G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, A. Z. Beaton, E. J. Benjamin, C. P. Benziger, A. Bonny, M. Brauer, M. Brodmann, T. J. Cahill, J. Carapetis, A. L. Catapano, S. S. Chugh, L. T. Cooper, J. Coresh, M. Criqui, N. DeCleene, K. A. Eagle, S. Emmons-Bell, V. L. Feigin, J. Fernández-Solà, G. Fowkes, E. Gakidou, S. M. Grundy, F. J. He, G. Howard, F. Hu, L. Inker, G. Karthikeyan, N. Kassebaum, W. Koroshetz, C. Lavie, D. Lloyd-Jones, H. S. Lu, A. Mirijello, A. M. Temesgen, A. Mokdad, A. E. Moran, P. Muntner, J. Narula, B. Neal, M. Ntsekhe, G. M. de Oliveira, C. Otto, M. Owolabi, M. Pratt, S. Rajagopalan, M. Reitsma, A. L. P. Ribeiro, N. Rigotti, A. Rodgers, C. Sable, S. Shakil, K. Sliwa-Hahnle, B. Stark, J. Sundström, P. Timpel, I. M. Tleyjeh, M. Valgimigli, T. Vos, P. K. Whelton, M. Yacoub, L. Zuhlke, C. Murray, V. Fuster, null null, G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, A. Beaton, E. J. Benjamin, C. P. Benziger, A. Bonny, M. Brauer, M. Brodmann, T. J. Cahill, J. R. Carapetis, A. L. Catapano, S. Chugh, L. T. Cooper, J. Coresh, M. H. Criqui, N. K. DeCleene, K. A. Eagle, S. Emmons-Bell, V. L. Feigin, J. Fernández-Sola, F. G. R. Fowkes, E. Gakidou, S. M. Grundy, F. J. He, G. Howard, F. Hu, L. Inker, G. Karthikeyan, N. J. Kassebaum, W. J. Koroshetz, C. Lavie, D. Lloyd-Jones, H. S. Lu, A. Mirijello, A. T. Misganaw, A. H. Mokdad, A. E. Moran, P. Muntner, J. Narula, B. Neal, M. Ntsekhe, G. M. Oliveira, C. M. Otto, M. O. Owolabi, M. Pratt, S. Rajagopalan, M. B. Reitsma, A. L. P. Ribeiro, N. A. Rigotti, A. Rodgers, C. A. Sable, S. S. Shakil, K. Sliwa, B. A. Stark, J. Sundström, P. Timpel, I. I. Tleyjeh, M. Valgimigli, T. Vos, P. K. Whelton, M. Yacoub, L. J. Zuhlke, M. Abbasi-Kangevari, A. Abdi, A. Abedi, V. Aboyans, W. A. Abrha, E. Abu-Gharbieh, A. I. Abushouk, D. Acharya, T. Adair, O. M. Adebayo, Z. Ademi, S. M. Advani, K. Afshari, A. Afshin, G. Agarwal, P. Agasthi, S. Ahmad, S. Ahmadi, M. B. Ahmed, B. Aji, Y. Akalu, W. Akande-Sholabi, A. Aklilu, C. J. Akunna, F. Alahdab, A. Al-Eyadhy, K. F. Alhabib, S. M. Alif, V. Alipour, S. M. Aljunid, F. Alla, A. Almasi-Hashiani, S. Almustanyir, R. M. Al-Raddadi, A. K. Amegah, S. Amini, A. Aminorroaya, H. Amu, D. A. Amugsi, R. Ancuceanu, D. Anderlini, T. Andrei, C. L. Andrei, A. Ansari-Moghaddam, Z. A. Anteneh, I. C. Antonazzo, B. Antony, R. Anwer, L. T. Appiah, J. Arabloo, J. Ärnlöv, K. D. Artanti, Z. Ataro, M. Ausloos, L. Avila-Burgos, A. T. Awan, M. A. Awoke, H. T. Ayele, M. A. Ayza, S. Azari, D. B. B, N. Baheiraei, A. A. Baig, A. Bakhtiari, M. Banach, P. C. Banik, E. A. Baptista, M. A. Barboza, L. Barua, S. Basu, N. Bedi, Y. Béjot, D. A. Bennett, I. M. Bensenor, A. E. Berman, Y. M. Bezabih, A. S. Bhagavathula, S. Bhaskar, K. Bhattacharyya, A. Bijani, B. Bikbov, M. M. Birhanu, A. Boloor, L. C. Brant,

H. Brenner, N. I. Briko, Z. A. Butt, F. L. C. dos Santos, L. E. Cahill, L. Cahuana-Hurtado, L. A. Cámera, I. R. Campos-Nonato, C. Cantu-Brito, J. Car, J. J. Carrero, F. Carvalho, C. A. Castañeda-Orjuela, F. Catalá-López, E. Cerin, J. Charan, V. K. Chattu, S. Chen, K. L. Chin, J.-Y. J. Choi, D.-T. Chu, S.-C. Chung, M. Cirillo, S. Coffey, S. Conti, V. M. Costa, D. K. Cundiff, O. Dadras, B. Dagnew, X. Dai, A. A. Damasceno, L. Dandona, R. Dandona, K. Davletov, V. D. la Cruz-Góngora, F. P. D. la Hoz, J.-W. D. Neve, E. Denova-Gutiérrez, M. D. Molla, B. T. Derseh, R. Desai, G. Deuschl, S. D. Dharmaratne, M. Dhimal, R. R. Dhungana, M. Dianatinasab, D. Diaz, S. Djalalinia, K. Dokova, A. Douiri, B. B. Duncan, A. R. Duraes, A. W. Eagan, S. Ebtehaj, A. Eftekhari, S. Eftekharzadeh, M. Ekholuenetale, N. E. Nahas, I. Y. Elgendy, M. Elhadi, S. I. El-Jaafary, S. Esteghamati, A. E. Etisso, O. Eyawo, I. Fadhil, E. J. A. Faraon, P. S. Faris, M. Farwati, F. Farzadfar, E. Fernandes, C. F. Prendes, P. Ferrara, I. Filip, F. Fischer, D. Flood, T. Fukumoto, M. M. Gad, S. Gaidhane, M. Ganji, J. Garg, A. K. Gebre, B. G. Gebregiorgis, K. Z. Gebregzabiher, G. G. Gebremeskel, L. Getacher, A. G. Obsa, A. Ghajar, A. Ghashghaee, N. Ghith, S. Giampaoli, S. A. Gilani, P. S. Gill, R. F. Gillum, E. V. Glushkova, E. V. Gnedovskaya, M. Golechha, K. B. Gonfa, A. H. Goudarzian, A. C. Goulart, J. S. Guadamuz, A. Guha, Y. Guo, R. Gupta, V. Hachinski, N. Hafezi-Nejad, T. G. Haile, R. R. Hamadeh, S. Hamidi, G. J. Hankey, A. Hargono, R. K. Hartono, M. Hashemian, A. Hashi, S. Hassan, H. Y. Hassen, R. J. Havmoeller, S. I. Hay, K. Hayat, G. Heidari, C. Herteliu, R. Holla, M. Hosseini, M. Hosseinzadeh, M. Hostiuc, S. Hostiuc, M. Househ, J. Huang, A. Humayun, I. Iavicoli, C. U. Ibeneme, S. E. Ibitoye, O. S. Ilesanmi, I. M. Ilic, M. D. Ilic, U. Iqbal, S. S. N. Irvani, S. M. S. Islam, R. M. Islam, H. Iso, M. Iwagami, V. Jain, T. Javaheri, S. K. Jayapal, S. Jayaram, R. Jayawardena, P. Jeemon, R. P. Jha, J. B. Jonas, J. Jonnagaddala, F. Joukar, J. J. Jozwiak, M. Jürisson, A. Kabir, T. Kahlon, R. Kalani, R. Kalhor, A. Kamath, I. Kamel, H. Kandel, A. Kandel, A. Karch, A. S. Kasa, P. D. Katoto, G. A. Kayode, Y. S. Khader, M. Khammarnia, M. S. Khan, M. N. Khan, M. Khan, E. A. Khan, K. Khatab, G. M. Kibria, Y. J. Kim, G. R. Kim, R. W. Kimokoti, S. Kisa, A. Kisa, M. Kivimäki, D. Kolte, A. Koolivand, V. A. Korshunov, S. L. K. Laxminarayana, A. Koyanagi, K. Krishan, V. Krishnamoorthy, B. K. Defo, B. K. Bicer, V. Kulkarni, G. A. Kumar, N. Kumar, O. P. Kurmi, D. Kusuma, G. F. Kwan, C. L. Vecchia, B. Lacey, T. Lallukka, Q. Lan, S. Lasrado, Z. S. Lassi, P. Lauriola, W. R. Lawrence, A. Laxmaiah, K. E. LeGrand, M.-C. Li, B. Li, S. Li, S. S. Lim, L.-L. Lim, H. Lin, Z. Lin, R.-T. Lin, X. Liu, A. D. Lopez, S. Lorkowski, P. A. Lotufo, A. Lugo, N. K. M, F. Madotto, M. Mahmoudi, A. Majeed, R. Malekzadeh, A. A. Malik, A. A. Mamun, N. Manafi, M. A. Mansournia, L. G. Mantovani, S. Martini, M. R. Mathur, G. Mazzaglia, S. Mehata, M. M. Mehndiratta, T. Meier, R. G. Menezes, A. Meretoja, T. Mestrovic, B. Miazgowski, T. Miazgowski, I. M. Michalek, T. R. Miller, E. M. Mirrakhimov, H. Mirzaei, B. Moazen, M. Moghadaszadeh, Y. Mohammad, D. K. Mohammad, S. Mohammed, M. A. Mohammed, Y. Mokhayeri, M. Molokhia, A. A. Montasir, G. Moradi, R. Moradzadeh, P. Moraga, L. Morawska, I. M. Velásquez, J. Morze, S. Mubarik, W. Muruet, K. I. Musa, A. J. Nagarajan, M. Nalini, V. Nangia, A. A. Naqvi, S. N. Swamy, B. R. Nascimento, V. C. Nayak, J. Nazari, M. Nazarzadeh, R. I. Negoi, S. N. Kandel, H. L. Nguyen, M. R. Nixon, B. Norrving, J. J. Noubiap, B. E. Nouthe, C. Nowak, O. O. Odukoya, F. A. Ogbo, A. T. Olagunju, H. Orru, A. Ortiz, S. M. Ostroff, J. R. Padubidri, R. Palladino, A. Pana, S. Panda-Jonas, U. Parekh, E.-C. Park, M. Parvizi, F. P. Kan, U. K. Patel, M. Pathak, R. Paudel, V. C. F. Pepito, A. Perianayagam,

N. Perico, H. Q. Pham, T. Pilgrim, M. A. Piradov, F. Pishgar, V. Podder, R. V. Polibin, A. Pourshams, D. R. Pribadi, N. Rabiee, M. Rabiee, A. Radfar, A. Rafiei, F. Rahim, V. Rahimi-Movaghar, M. H. U. Rahman, M. A. Rahman, A. M. Rahmani, I. Rakovac, P. Ram, S. Ramalingam, J. Rana, P. Ranasinghe, S. J. Rao, P. Rathi, L. Rawal, W. F. Rawasia, R. Rawassizadeh, G. Remuzzi, A. M. Renzaho, A. Rezapour, S. M. Riahi, R. L. Roberts-Thomson, L. Roever, P. Rohloff, M. Romoli, G. Roshandel, G. M. Rwegerera, S. Saadatagah, M. M. Saber-Ayad, S. Sabour, S. Sacco, M. Sadeghi, S. S. Moghaddam, S. Safari, A. Sahebkar, S. Salehi, H. Salimzadeh, M. Samaei, A. M. Samy, I. S. Santos, M. M. Santric-Milicevic, N. Sarrafzadegan, A. Sarveazad, T. Sathish, M. Sawhney, M. Saylan, M. I. Schmidt, A. E. Schutte, S. Senthilkumaran, S. G. Sepanlou, F. Sha, S. Shahabi, I. Shahid, M. A. Shaikh, M. Shamali, M. Shamsizadeh, M. S. R. Shawon, A. Sheikh, M. Shigematsu, M.-J. Shin, J. I. Shin, R. Shiri, I. Shiue, K. Shuval, S. Siabani, T. J. Siddiqi, D. A. Silva, J. A. Singh, A. S. Mtech, V. Y. Skryabin, A. A. Skryabina, A. Soheili, E. E. Spurlock, L. Stockfelt, S. Stortecky, S. Stranges, R. S. Abdulkader, H. Tadbiri, E. G. Tadesse, D. B. Tadesse, M. Tajdini, M. Tariqujjaman, B. F. Teklehaimanot, M.-H. Temsah, A. K. Tesema, B. Thakur, K. R. Thankappan, R. Thapar, A. G. Thrift, B. Timalsina, M. Tonelli, M. Touvier, M. R. Tovani-Palone, A. Tripathi, J. P. Tripathy, T. C. Truelsen, G. M. Tsegay, G. W. Tsegaye, N. Tsilimparis, B. S. Tusa, S. Tyrovolas, K. K. Umapathi, B. Unim, B. Unnikrishnan, M. S. Usman, M. Vaduganathan, P. R. Valdez, T. J. Vasankari, D. Z. Velazquez, N. Venketasubramanian, G. T. Vu, I. S. Vujcic, Y. Waheed, Y. Wang, F. Wang, J. Wei, R. G. Weintraub, A. H. Weldemariam, R. Westerman, A. S. Winkler, C. S. Wiysonge, C. D. Wolfe, B. L. Wubishet, G. Xu, A. Yadollahpour, K. Yamagishi, L. L. Yan, S. Yandrapalli, Y. Yano, H. Yatsuya, T. Y. Yeheyis, Y. Yeshaw, C. S. Yilgwan, N. Yonemoto, C. Yu, H. Yusefzadeh, G. Zachariah, S. B. Zaman, M. S. Zaman, M. Zamanian, R. Zand, A. Zandifar, A. Zarghi, M. S. Zastrozhin, A. Zastrozhina, Z.-J. Zhang, Y. Zhang, W. Zhang, C. Zhong, Z. Zou, Y. M. H. Zuniga, C. J. Murray, and V. Fuster. "Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019". *Journal of the American College of Cardiology*, Vol. 76, No. 25, pp. 2982–3021, 2020.

[Rumb 03] J. A. Rumberger and L. Kaufman. "A rosetta stone for coronary calcium risk stratification: agatston, volume, and mass scores in 11,490 individuals". *American Journal of Roentgenology*, Vol. 181, No. 3, pp. 743–748, 2003.

[Rume 86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". *Nature*, Vol. 323, No. 6088, pp. 533–536, 1986.

[Russ 15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. "Imagenet large scale visual recognition challenge". *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252, 2015.

[Shri 12] J. E. Shriki, J. S. Shinbane, M. A. Rashid, A. Hindoyan, J. G. Withey, A. DeFrance, M. Cunningham, G. R. Oliveira, B. H. Warren, and A. Wilcox. "Identifying, characterizing, and classifying congenital anomalies of the coronary arteries". *Radiographics*, Vol. 32, No. 2, pp. 453–468, 2012.

[Taub 20] O. Taubmann, J. Li, F. Denzinger, E. Eibenberger, F. C. Müller, M. W. Brejnebøl, and A. Maier. "Automatic detection of free intra-abdominal air in computed tomography". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 232–241, Springer, 2020.

[Tayl 13] C. A. Taylor, T. A. Fonte, and J. K. Min. "Computational fluid dynamics applied to cardiac computed tomography for noninvasive quantification of fractional flow reserve: scientific basis". *Journal of the American College of Cardiology*, Vol. 61, No. 22, pp. 2233–2241, 2013.

[Tham 21] F. Thamm, O. Taubmann, F. Denzinger, M. Jürgens, H. Ditt, and A. Maier. "SyNCCT: Synthetic Non-contrast Images of the Brain from Single-Energy Computed Tomography Angiography". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 681–690, Springer, 2021.

[Tham 22] F. Thamm, O. Taubmann, M. Jürgens, A. Thamm, F. Denzinger, L. Rist, H. Ditt, and A. Maier. "Building Brains: Subvolume Recombination for Data Augmentation in Large Vessel Occlusion Detection". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 634–643, Springer, 2022.

[Tion 05] A. Y. Tiong and D. Brieger. "Inflammation and coronary artery disease". *American Heart Journal*, Vol. 150, No. 1, pp. 11–18, 2005.

[Van 17] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts. "Computational radiomics system to decode the radiographic phenotype". *Cancer research*, Vol. 77, No. 21, pp. e104–e107, 2017.

[Van 64] R. Van Praagh, S. Van Praagh, P. Vlad, and J. D. Keith. "Anatomic types of congenital dextrocardia: diagnostic and embryologic implications". *American Journal of Cardiology*, Vol. 13, No. 4, pp. 510–531, 1964.

[Viti 22] M. Viti, H. Talbot, and N. Gogin. "Transformer Graph Network for Coronary Plaque Localization in CCTA". In: *2022 IEEE 19th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2022.

[Wagn 22] F. Wagner, M. Thies, F. Denzinger, M. Gu, M. Patwari, S. Ploner, N. Maul, L. Pfaff, Y. Huang, and A. Maier. "Trainable joint bilateral filters for enhanced prediction stability in low-dose CT". *Scientific Reports*, Vol. 12, No. 1, pp. 1–9, 2022.

[Wagn 23] F. Wagner, M. Thies, L. Pfaff, O. Aust, S. Pechmann, D. Weidner, N. Maul, M. Rohleder, M. Gu, J. Utz, F. Denzinger, and A. Maier. "On the Benefit of Dual-domain Denoising in a Self-supervised Low-dose CT Setting". In: *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, IEEE, 2023.

[Whit 21] R. D. White, B. S. Erdal, M. Demirer, V. Gupta, M. T. Bigelow, E. Dikici, S. Candemir, M. S. Galizia, J. L. Carpenter, T. P. O'Donnell, *et al.* "Artificial intelligence to assist in exclusion of coronary atherosclerosis during CCTA evaluation of chest pain in the emergency department: preparing an application for real-world use". *Journal of Digital Imaging*, Vol. 34, pp. 554–571, 2021.

[Yang 21] X. Yang, T. Han, R. Kang, J. Fan, and D. Ai. "Automatic Localization and Classification of Coronary Artery Plaques from Cardiac CTA with A Boundary-Constrained 3D Fully Convolutional Network". In: *2021 5th International Conference on Advances in Image Processing*, pp. 90–96, 2021.

[Zbon 21] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. "Barlow twins: Self-supervised learning via redundancy reduction". In: *International Conference on Machine Learning*, pp. 12310–12320, PMLR, 2021.

[Zhan 22] Y. Zhang, J. Ma, and J. Li. "Coronary R-CNN: Vessel-Wise Method for Coronary Artery Lesion Detection and Analysis in Coronary CT Angiography". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 207–216, Springer, 2022.

[Zhen 13] Y. Zheng, H. Tek, and G. Funka-Lea. "Robust and accurate coronary artery centerline extraction in CTA by combining model-driven and data-driven approaches". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 74–81, Springer, 2013.

[Zrei 18] M. Zreik, R. W. Van Hamersvelt, J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum. "A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography". *IEEE Transactions on Medical Imaging*, Vol. 38, No. 7, pp. 1588–1598, 2018.

[Zrei 19] M. Zreik, R. W. van Hamersvelt, N. Khalili, J. M. Wolterink, M. Voskuil, M. A. Viergever, T. Leiner, and I. Išgum. "Deep learning analysis of coronary arteries in cardiac CT angiography for detection of patients requiring invasive coronary angiography". *IEEE Transactions on Medical Imaging*, Vol. 39, No. 5, pp. 1545–1557, 2019.