Machine learning based analysis of cardiovascular images

Jelmer M. Wolterink

This thesis was typeset by the author using LaTeX 2_ε
ISBN: 978-94-6299-587-1
Cover design: Jochem Kruizinga
Layout: Jelmer Wolterink
Printed by: Ridderprint BV

Machine learning based analysis of cardiovascular images

Analyse van cardiovasculaire beelden
op basis van machine learning
(met een samenvatting in het Nederlands)

Proefschrift
ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op donderdag 11 mei 2017 des middags te 2.30 uur

door

Jelmer Maarten Wolterink
geboren op 13 augustus 1988
te Den Ham

Promotoren: Prof. dr. ir. M.A. Viergever
Prof. dr. T. Leiner

Copromotor: Dr. I. Išgum

De publicatie van dit proefschrift werd mogelijk gemaakt met financiële steun van Philips Healthcare en Pie Medical Imaging.

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

Contents

Chapter 1

Introduction

1.1 Cardiovascular diseases

Cardiovascular diseases (CVDs) are a collection of disorders that affect the heart and blood vessels. The World Health Organization estimates that 17.5 million deaths in 2012 were due to CVDs, of which 7.4 million caused by coronary artery disease (CAD), and 10.1 million caused by other CVDs, including stroke and congenital heart defects (CHD) [1]. This accounts for 31% of all global deaths, making CVDs the world's leading cause of death.

1.1.1 CT imaging of coronary artery disease

In patients with CAD, atherosclerotic plaque lesions develop in the coronary artery wall [2]. Plaque formation can follow different mechanisms, depending on location and individual risk factors. However, the process typically starts with the accumulation of lipid-rich material in the intimal vessel wall. Over an extended period of time, which may be years or decades, other cells infiltrate the lesion and form a necrotic core. The tissue separating the plaque from the lumen is gradually replaced by a thin fibrous cap, and the necrotic core can calcify over time, forming coronary artery calcification (CAC) [3].

Atherosclerotic lesions in the coronary arteries can have two adverse effects on the oxygen supply to the myocardium. First, plaque build-up may cause a narrowing in the coronary artery. This can trigger angina and is often reason for percutaneous coronary interventions or coronary artery bypass grafts. Second, the thin cap separating vulnerable plaques from the lumen may rupture. This causes the formation of a blood clot that obstructs the blood flow to the myocardium. If a patient is not immediately treated with antithrombotic agents or an intervention, this may lead to myocardial infarction and death.

In around one third of patients, sudden cardiac death due to myocardial infarction is the first manifestation of CAD [4]. Timely identification of these patients could lead to the prevention of coronary events. To this end, a wide range of factors has been recognized, including non-modifiable risk factors such as sex, age and family history, and modifiable risk factors such as LDL cholesterol, hypertension and smoking status [3].

In addition to these traditional risk factors indicating risk of CAD events, medical imaging has been used to identify patients at risk. Cardiac computed tomography (CT) provides a high-resolution non-invasive view of the coronary arteries and can thus be used to identify atherosclerotic plaques. In a typical cardiac CT examination one scan is acquired before injection of a contrast agent and one scan after injection. While the scan acquired after contrast injection provides excellent visualization of the coronary arteries, the non-contrast-enhanced scan can be used to quantify coronary artery calcification. The presence of CAC does not signal the location of vulnerable atherosclerotic lesions [5, 6], but the total amount of CAC as

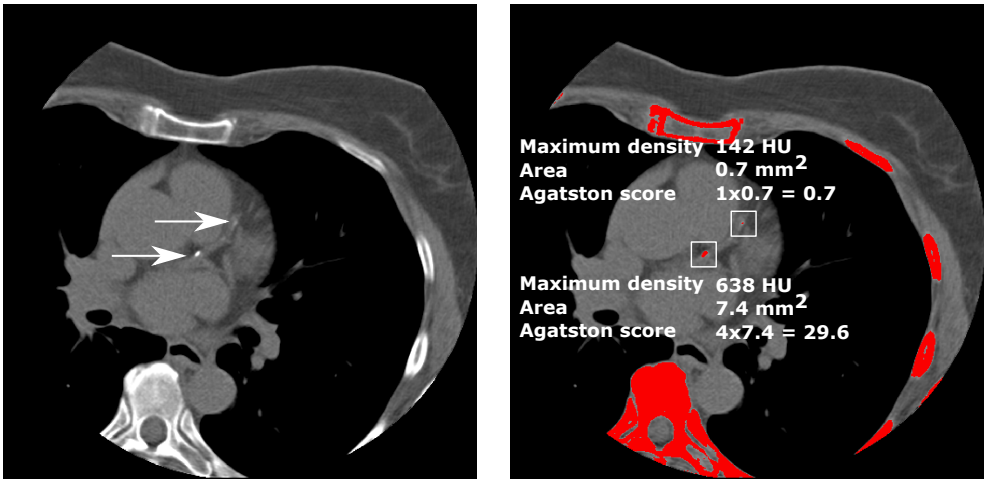


Figure 1.1: Coronary artery calcium scoring in non-contrast-enhanced cardiac CT. The image shows two calcified lesions (≥ 130 HU) in the left anterior descending coronary artery, indicated by white arrows. The maximum density in each lesion determines a weight factor (1: 130-200 HU, 2: 200-300 HU, 3: 300-400 HU, 4: >400 HU). The Agatston score of the lesion in this slice is computed as the product of the weight factor and the lesion area.

quantified in cardiac CT is a strong predictor of CVD events, independently or in combination with traditional risk factors [7, 8].

To determine the amount of CAC in a non-contrast-enhanced cardiac CT image, lesions consisting of groups of voxels above a threshold of 130 Hounsfield units (HU) in the coronary arteries are identified. Each CAC lesion is assigned a score using its area and maximum in-plane attenuation (Fig. 1.1). The sum of these lesion-based scores determines the patient's CAC score, or Agatston score [9]. The Agatston score can be used to assign patients to a cardiovascular risk category, ranging from very low CVD risk (Agatston score 0) to very high CVD risk (Agatston score > 400) [10, 7]. Although alternative CAC quantification methods based on calcium volume or mass have been proposed [11], the Agatston score is the most widely clinically used method for CAC quantification.

In current clinical practice, CAC is scored by manual identification of each lesion. This is a time-consuming and tedious process when performed in a large number of scans. Recent guidelines on the assessment of CVD risk recommend calcium scoring in adults at intermediate and low-to-intermediate risk according to traditional risk factors [12]. This is likely to further increase the number of CAC scoring examinations. Therefore, (semi-)automatic methods have been proposed to relieve the burden for clinical experts [13, 14, 15, 16]. Chapter 2 of this thesis describes a method that fully automatically identifies CAC lesions, labels them according to their location in the coronary tree, and assigns the patient to a CVD risk category. Furthermore, the method allows an expert to assess and optionally correct lesion

predictions with minimal effort. This is beneficial in the case of CVD patient analysis, where a higher accuracy at the level of individual lesions may be required. In order to assess the clinical potential of automatic CAC scoring methods, a detailed comparison of their performance is provided in Chapter 3 of this thesis, using a representative data set and standardized evaluation criteria.

CT provides excellent visualization of the coronary arteries, but its ionizing radiation may carry health risks [17, 18]. Therefore, recent years have seen a trend towards radiation dose reduction in CT examinations. Rigorous dose reduction may be obtained by omitting CT acquisitions altogether. It has been shown that there is a strong correlation between the amount of CAC in non-contrast-enhanced CT and contrast-enhanced CT images [19, 20]. Hence, automatic CAC quantification in contrast-enhanced CT, as described in Chapter 5 of this thesis, could obviate the need for a dedicated non-contrast-enhanced CT scan. Alternatively, CT images may be acquired with reduced tube voltage or tube current, and methods such as proposed in Chapter 7 could be used to limit the loss in image quality resulting from dose reduction.

1.1.2 MR imaging of congenital heart disease

Patients with congenital heart disease (CHD) are born with a spectrum of malformations of the heart and vasculature. This affects almost 1% of live births and often necessitates surgery during childhood [21]. Due to the low age of most CHD patients, pre-operative imaging is typically performed using magnetic resonance imaging (MRI). MRI does not require ionizing radiation and thus carries fewer risks for the patient. It has been shown that 3D printed models based on cardiovascular MR imaging are useful for pre-operative planning [22]. However, obtaining such a 3D model based on an MR image is difficult due to large anatomical differences among patients and non-standardized image acquisition protocols.

Manual segmentation of cardiac MR images in patients with CHD can require up to several hours [23]. Therefore, (semi-)automatic methods that perform segmentation of MR images with high accuracy can substantially reduce the workload for clinicians. Chapter 8 describes a method to automatically obtain a segmentation of the blood pool and myocardium from a cardiovascular MR image, which may then be used to obtain a patient-specific 3D model.

1.2 Machine learning

This thesis describes the analysis of cardiac CT and MR images using machine learning techniques. Such techniques train models to map an input to an output value, based on information extracted from the input. A model may either perform regression, where a continuous value is predicted, or classification, in which the input sample is assigned to a discrete class.

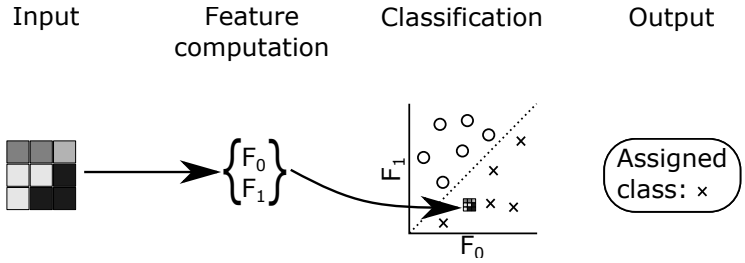


Figure 1.2: A conventional machine learning pipeline. Predefined features are computed for the sample input. These features are used to classify the sample based on previously seen samples.

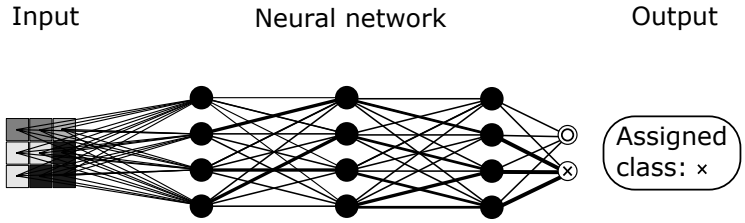


Figure 1.3: A deep learning pipeline. An artificial neural network transforms the input into features, and subsequently into an output. Network weights have previously been optimized during training.

Fig. 1.2 shows a conventional machine learning pipeline for classification. To classify an input sample, variables or features that describe the input are first computed. The classifier has previously seen a number of representative samples and their labels during a *training* stage. When *testing* the new sample, it is placed in a feature space and classified according to the class of the training samples to which its feature values are most similar.

The exact representation of samples, features and outputs depends on the specific medical imaging problem. For example, in the case of coronary calcium scoring (Chapter 2) samples are candidate calcifications, in MRI segmentation they are voxels (Chapter 8), and in organ localization (Chapter 4) samples are 2D image slices. Likewise, the design of features depends on the problem, and is often based on expert domain knowledge and expertise.

In deep learning techniques, the explicit design of features is omitted [22, 24]. Instead, an artificial neural network (ANN) connects the input and output through a number of layers with weighted connections (Fig. 1.3). To train the ANN, pairs of inputs and outputs are presented to the ANN. Each time such a pair is shown, the weights on the connections of the ANN are updated so that the error with respect to the target output is minimized. After a sufficient number of iterations, the ANN is able to classify a new and unseen sample. In image processing, the layers of an ANN typically consist of trainable convolution kernels, which form a convolutional neural

network (CNN) [25]. Activation patterns at the inner layers of a CNN represent features that are used to classify the input in and end-to-end fashion [26].

Over the past few years, deep learning methods have achieved state-of-the-art results in a wide range of applications. In medical image analysis, this approach has achieved excellent results in applications that previously required domain-specific feature design, e.g. brain segmentation [27] or pancreas segmentation [28]. Moreover, it has been shown that features extracted from one data set can be powerful for classification in different domains [29, 30, 31]. Chapters 4-9 of this thesis describe image analysis techniques that were developed based on deep learning methods.

1.3 Thesis outline

This thesis presents machine learning methods for analysis of CT images of patients with coronary artery disease and MR images of patients with congenital heart disease.

Chapter 2 describes a randomized forest-based method for automatic coronary calcium scoring in non-contrast-enhanced cardiac CT scans. The method identifies, labels and quantifies coronary calcifications and optionally presents lesions with ambiguous classification results to an expert for review.

Chapter 3 describes an evaluation framework for (semi-)automatic methods for coronary artery calcium scoring in non-contrast-enhanced cardiac CT scans, in which a diverse data set and standardized evaluation criteria are provided.

Chapter 4 describes a method for 3D organ localization in medical imaging volumes based on 2D classification of axial, sagittal and coronal slices using a convolutional neural network.

Chapter 5 describes a machine learning method for automatic coronary calcium scoring in contrast-enhanced cardiac CT scans using a pair of convolutional neural networks, one to identify candidate voxels, and one to classify the selected voxels.

Chapter 6 describes a machine learning method for coronary centerline extraction from contrast-enhanced cardiac CT scans. The method iteratively determines the most likely direction and radius of the coronary artery based on a combined convolutional neural network performing classification and regression.

Chapter 7 describes a deep convolutional generative adversarial network for noise reduction in low-dose CT scans.

Chapter 8 describes a dilated convolutional neural network for segmentation of myocardium and blood pool in cardiovascular MR images of patients with congenital heart disease.

Chapter 9 describes a series of experiments in which a single convolutional neural network is trained to perform coronary artery segmentation in cardiac CT, brain tissue segmentation in brain MRI and pectoral muscle segmentation in breast MRI.

Chapter 10 summarizes the work presented in this thesis and discusses future directions.

Chapter 2

Automatic coronary calcium scoring in non-contrast-enhanced ECG-triggered cardiac CT with ambiguity detection

Based on:
J.M. Wolterink, T. Leiner, R.A.P. Takx, M.A. Viergever, I. Išgum. "Automatic Coronary Calcium Scoring in Non-Contrast-Enhanced ECG-Triggered Cardiac CT with Ambiguity Detection," *IEEE Transactions on Medical Imaging*, 2015, vol. 34, pp. 1867–1878

Abstract

The amount of coronary artery calcification (CAC) is a strong and independent predictor of cardiovascular events. We present a system that automatically quantifies total patient and per coronary artery CAC in non-contrast-enhanced, ECG-triggered cardiac CT. The system identifies candidate calcifications that cannot be automatically labeled with high certainty and optionally presents these to an expert for review.

Candidates were extracted by intensity-based thresholding and described by location features derived from estimated coronary artery positions, as well as size, shape and intensity features. Next, a two-class classifier distinguished between coronary calcifications and negatives or a multiclass classifier labeled CAC per coronary artery. Candidates that could not be labeled with high certainty were identified by entropy-based ambiguity detection and presented to an expert for review and possible relabeling.

The system was evaluated with 530 test images. Using the two-class classifier, the intraclass correlation coefficient (ICC) between reference and automatically determined total patient CAC volume was 0.95. Using the multiclass classifier, the ICC between reference and automatically determined per artery CAC volume was 0.98 (LAD), 0.69 (LCX), and 0.95 (RCA). In 49% of CTs, no ambiguous candidates were identified, while review of the remaining CTs increased the ICC for total patient CAC volume to 1.00, and per artery CAC volume to 1.00 (LAD), 0.95 (LCX), and 0.99 (RCA).

In conclusion, CAC can be automatically identified in non-contrast-enhanced ECG-triggered cardiac CT. Ambiguity detection with expert review may enable the application of automatic CAC scoring in the clinic with a performance comparable to that of a human expert.

2.1 Introduction

Cardiovascular disease (CVD) is the global leading cause of death [1]. The amount of coronary artery calcification (CAC) has been shown to be a strong and independent predictor of CVD events, such as myocardial infarction [7, 8]. Hence, to estimate the risk of a CVD event, CAC is routinely quantified in non-contrast-enhanced ECG-triggered calcium scoring CT (CSCT). In current clinical practice, calcified plaque in the coronary arteries is manually identified using commercially available semi-automatic software. To label coronary artery calcifications, typically all voxels above a standard threshold of 130 HU are considered [9]. An expert subsequently identifies voxels that represent coronary calcifications and labels them according to the coronary artery they are located in. Identified calcifications are quantified using e.g. an Agatston [9] or volume score [32]. These scores indicate a patient's risk of CVD [7, 8].

Although manual CAC identification is not very complicated, it may be time-consuming and impractical in clinical practice, large studies and possible screening settings. Automatic computer-aided identification of CAC would decrease the workload of experts.

A number of automatic and semi-automatic algorithms for the identification of CAC in non-contrast-enhanced CT have been proposed. Several of these algorithms have identified coronary calcifications among a set of high-intensity 3D candidate calcifications, using supervised k-nearest neighbor (k-NN) or support vector machine (SVM) classifiers. The candidate calcifications to be classified were described by location, size, shape and intensity characteristics. Particularly, location characteristics have been shown to be indispensable for accurate identification. However, owing to poor contrast between coronary arteries and surrounding tissues, the computation of location characteristics in CSCT is very challenging. Therefore, several algorithms have described candidate calcifications by their spatial relation to cardiac structures. Išgum et al. automatically segmented the aorta and the heart and described each candidate's location with respect to these segmentations [33]. Kurkure et al. and Brunner et al. described the location of candidates using a heart-centered coordinate system, determined by automatic heart segmentation [34] and user-defined seed points [35]. Recently, Shahzad et al. proposed an estimation of the position of coronary arteries [13]. Except for the method described in [13], which required a large set of coronary CT angiography (CCTA) scans, these methods only required CSCT scans for the computation of location characteristics. A different approach was proposed by Saur et al. [36], where CSCT scans were combined with CCTA scans for segmentation of the aorta and the coronary arteries, followed by rule-based detection of calcifications in CSCT. Recently, algorithms for the automatic identification of CAC in non-ECG-triggered non-contrast-enhanced chest CT have been proposed. For this purpose, Išgum et al. designed a map providing a priori probabilities for the spatial appearance of CAC [14]. Location characteristics were derived from this map, and together with size and intensity char-

acteristics used to identify CAC in a supervised classification approach similar to [33, 35, 34, 13]. Xie et al. identified CAC as connected groups of high-intensity voxels in a heart region constrained by segmentations of the lungs, bone, aorta and fatty tissue [37]. These two algorithms analyzed chest scans acquired with multidetector CT, while all other algorithms analyzed cardiac scans acquired using electron-beam [34, 35], multidetector [33, 13] and dual-source [36] CT. The proposed methods identified calcifications irrespective of the coronary artery they were located in, except [35] and [13], where per artery CAC burden was determined in addition to the overall calcium score.

Optimal CVD risk stratification or monitoring of CAC progression over time in individual patients[38] requires high accuracy of automatically determined scores, likely at the level of interobserver agreement [39]. The performance of existing methods is still below this level. Previously proposed methods are error-prone at typical and challenging locations, e.g. at the boundary between coronary arteries and the ascending aorta, or they incorrectly identify voxels with high noise levels as calcifications.

Here, we propose a system for the automatic identification and quantification of total patient and per artery CAC burden in routinely acquired CSCT images. Namely, the system identified coronary calcifications by means of either a two-class classifier that distinguished between coronary calcifications and negatives, or a multiclass classifier that labeled coronary calcifications according to the coronary artery they were located in. Like in previous work, candidates were high-intensity lesions. These were classified based on size, shape, intensity and location features. To compute location features, the position of the coronary arteries was estimated using a small set of CCTA images with manually delineated coronary arteries. To enable application of the proposed system in clinical settings, candidates that could not be labeled with high certainty were automatically detected. Only these candidates were presented to an expert for inspection and - if needed - relabeling. The preferred degree of expert interaction was determined through ambiguity detection, based on a desired accuracy level and time constraints, i.e. the number of candidates or scans that may be reviewed. A preliminary version of this algorithm has been described in [40]. This chapter extends the preliminary work in several ways. First, besides detection of CAC as one class, detection of CAC per coronary artery is proposed. Consequently, the ambiguity detection mechanism is modified to provide multiclass ambiguity detection. Second, for faster estimation of the location of the coronary arteries the algorithm has been modified to use non-contrast-enhanced cardiac CT scans as atlases. Next, to demonstrate the robustness of the algorithm, the data set has been substantially extended. Finally, the evaluation has been extended to demonstrate the performance on CT images acquired using CT scanners of different vendors.

2.2 Data

This study retrospectively included 1013 consecutive CT scans that were routinely acquired for clinical identification and quantification of CAC (CAC scoring) at the University Medical Center Utrecht, The Netherlands. The need for informed consent was waived by the local medical ethics committee. Scans were made between August 2011 and May 2013 with a 256-detector row Philips Brilliance iCT scanner (tube voltage 120 kVp, tube current 55 mAs) during a single breath-hold, with ECG-triggering and without contrast enhancement. Scan duration was approximately 5 s. The images were reconstructed to 3 mm section thickness with in-plane resolution ranging from 0.29 mm to 0.49 mm, depending on patient size. Scans of patients below the age of 18 (18) were excluded from this study. These young patients were scanned because of a history of congenital heart disease, resulting in an unusual anatomy. Typical calcified atherosclerotic plaques are not likely to be found in these patients [41]. Furthermore, scans of patients with stents or metal implants causing beam hardening artifacts (81) were excluded. In the remaining 914 scans, patient age ranged from 18 to 88 (median 52) and 280 patients (31%) were female.

To define a reference standard, coronary artery calcifications in the left main (LM), left anterior descending (LAD), left circumflex (LCX) and right coronary artery (RCA) were manually identified by experts using commercially available software (HeartBeat CS, Philips, Cleveland, Ohio), as part of clinical routine. Because it might be difficult to visually distinguish between calcifications at the border of the LM and the LAD, in this work, both were considered LAD calcifications. A standard calcium scoring threshold of 130 HU was used.

To compare the performance of the automatic system with that of a second expert, a subset of 156 consecutive scans was manually annotated by a physician with five years of experience (>1000 scans) in CSCTs.

Finally, for ten consecutive CSCT scans that did not contain any coronary calcifications or anatomical abnormalities, corresponding CCTA scans were included. These CCTA scans were used to estimate the position of the coronary arteries. Each CSCT and its corresponding CCTA were acquired in a single session with the same scanner. CCTA scans were acquired during a single breath-hold, tube voltage was 120 kVp and tube current varied from 210 to 300 mAs. Scan duration was approximately 5 s. Images were reconstructed to 0.9 mm section thickness and in-plane resolution equal to the CSCT images. LAD, LCX and RCA centerlines were manually delineated in the CCTA scans using custom software. The LM was considered to be part of the LAD and diagonal, septal or marginal branches were not annotated.

2.3 Automatic coronary calcium scoring and ambiguity detection

To automatically identify and quantify calcifications in the coronary arteries, a supervised machine learning system was developed. First, candidates for CAC were extracted from the images based on their intensity. Each of these candidate calcifi-

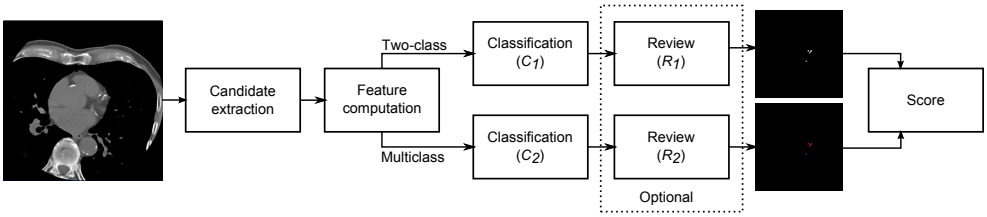


Figure 2.1: Overview of the proposed system. A CSCT scan was the input to the system. First, candidates for CAC were extracted from the input image. Next, each candidate was described using a set of size, shape, intensity and location features. Using either two-class (C_1, CAC vs. negatives) or multiclass (C_2, LAD CAC vs. LCX CAC vs. RCA CAC vs. negatives) classification, CAC lesions were identified among the candidates. An optional review step allowed two-class (R_1) or multiclass (R_2) relabeling of lesions which the system could not label with high certainty. Finally, identified CAC was quantified with the volume and Agatston score.

cations was described with a set of features, characterizing its size, shape, intensity and location information. Candidate calcifications either belonged to the class of calcifications in the LAD (y_{LAD}), LCX (y_{LCX}) or RCA (y_{RCA}), or to the class of negatives (y_{neg}). They were classified by means of a two-class ($y_{LAD} \cup y_{LCX} \cup y_{RCA}$ vs. y_{neg}) or multiclass (y_{LAD} vs. y_{LCX} vs. y_{RCA} vs. y_{neg}) strategy. Potential calcifications that the system could not label with high certainty were detected and optionally selected for expert review. Finally, identified CAC was quantified with the volume and Agatston score. Fig. 2.1 shows an overview of this system.

2.3.1 Candidate extraction

In clinical practice, components of connected voxels inside the coronary artery wall with intensity above 130 HU are considered to be CAC. Hence, 3D connected components, defined using 26-connectivity, which were above the intensity threshold value of 130 HU were considered candidates for CAC [14]. These candidates represented CAC, calcifications in other arteries (e.g. aorta), other cardiac calcifications (e.g. aortic valve and mitral valve), bony structures and noise. To prevent erroneous identification of noise voxels as CAC, candidates smaller than 1.5 mm^3 were not considered as candidates for CAC. Similarly, candidates larger than 1500 mm^3 were excluded as they likely represented bony structures.

2.3.2 Coronary centerline estimation

Previous work on the identification of CAC in CSCT emphasized the importance of location information [14]. Accurate segmentation of the coronary arteries would substantially simplify the identification of CAC. However, this is very challenging, if at all possible, due to low intensity gradients between blood and soft tissue in CSCT. In the algorithm proposed here, patient-specific centerlines were independently es-

timated for three major coronary arteries (LAD, LCX and RCA), allowing a detailed positioning of a candidate's location with respect to all three major coronary arteries.

The position of each coronary artery centerline in a patient CT was estimated using a set of ten selected CSCT atlases with defined centerlines. Poor visibility of the coronary arteries makes manual delineation of the centerlines in CSCT challenging. Therefore, the centerlines were manually annotated in a corresponding CCTA image, and subsequently propagated from the CCTA to the CSCT. Thereafter, the centerlines were propagated from the CSCT atlas to a CSCT patient image. To propagate the centerlines, scans were aligned using affine and subsequent elastic registration, with *elastix* [42]. An iterative stochastic gradient descent optimizer was used for optimization of the cost function (negative mutual information). A detailed description of registration parameter settings is given in [14].

Alignment with ten atlases resulted in ten estimates of each centerline C. The most likely estimate $\hat{C}$ of C was determined iteratively as a combination of propagated centerlines, in line with the algorithm described by Langerak et al. [43]. Initially, all propagated centerlines were selected and an estimation $\hat{C}_0$ was defined as their geometric median [44], [45]. In each iteration i, the single propagated centerline that had the largest discrete Fréchet distance to $\hat{C}_i$ was discarded [46], and a new estimate $\hat{C}_{i+1}$ was determined as the geometric median of the remaining centerlines. The algorithm terminated when either the discrete Fréchet distance of all propagated centerlines to $\hat{C}_i$ was less than a predefined maximum (20 mm), or three propagated centerlines remained. The geometric median of the remaining centerlines was the final estimate $\hat{C}$.

2.3.3 Feature computation

Previous work demonstrated that coronary calcifications can be distinguished from other candidates using features describing candidates' size, shape, intensity, and location characteristics [33, 34, 35, 14, 37, 13]. The proposed method used a selection of these features, supplemented with location features specific to the estimated coronary artery centerlines.

A candidate's size was described by its volume in mm^3. Shape features distinguished between plate-like (e.g. mitral and aortic valve) and elongated (e.g. large coronary) candidate calcifications [33, 13]. These were calculated as the ratios $\frac{\lambda_1}{\lambda_3}$ and $\frac{\lambda_2}{\lambda_3}$, where $\lambda_1 \leq \lambda_2 \leq \lambda_3$ are eigenvalues found using a principal component analysis of the voxels comprising each candidate. Gaussian filters from 0^{th} through 2^{nd} order in three image directions (x, y and z) at five scales were computed at the center of gravity of each candidate. Scales of 0.3, 0.6, 1.2, 2.4, 4.8 mm were chosen, with the smallest scale matching the highest in-plane resolution among all images. In addition, each candidate was described by the maximum, average and standard deviation of the intensity values among its voxels.

A set of location features related the position of candidates to the three estimated

coronary artery centerlines (LAD, LCX, RCA). For each candidate the minimum, average and maximum Euclidean distance among its voxels to each centerline were computed. In addition, the minimum Euclidean distances obtained in this way were used to compute the minimum distance to *any* coronary artery as well as the average distance to *all* coronary arteries. Coronary calcifications appear more frequently in proximal than in distal parts of the arteries. Hence, for each candidate the closest point on each of the estimated centerlines was found, and the arc lengths from those points to the origins of the centerlines were used as features. In addition, the location of candidates within the heart was determined. Ideally, this location would be determined with respect to anatomical structures such as the apex of the heart or mitral valve insertion points. However, the limited contrast in CSCT makes it very challenging to automatically identify these structures. Hence, a heart-center point was determined as the average of the starting points of the estimated LAD and RCA, and the Euclidean distance of each candidate to this point was computed [13]. All distances were expressed in mm.

A candidate's location in an image coordinate space has been shown to provide valuable information [14], but large variations in anatomical coverage of cardiac CT scans prohibit the direct use of the patient CT image space. Therefore, the location of all candidates was expressed in the coordinate space of a single atlas. Although artery estimation (Section 2.3.2) selected atlases which aligned well with the patient CT, their coordinate systems were not used for two reasons. First, doing so would introduce a bias towards one particular atlas. Second, none of these atlases would necessarily be selected for all patients. Therefore, an additional independent atlas A_I was utilized. To exploit the information provided by artery estimation, each candidate's coordinates were first propagated to the M atlases selected for artery estimation and thereafter, to the independent atlas A_I. This resulted in M locations in A_I. These locations were averaged to provide three features, the x, y and z coordinates in the coordinate space of A_I.

Noise voxels in cardiac CT scans often appear clustered at typical locations (e.g. at the level of the apex), while other candidates, including CAC, have a sparser distribution. To differentiate between these, the average Euclidean distance (in mm) between the centers of gravity of the candidate and the nearest three other candidates in the image was computed.

The total set consisted of 76 features: 1 size, 2 shape, 53 intensity and 20 location features.

2.3.4 Classification

Two supervised classifiers were trained to assign each candidate to a class $y \in Y$ based on its feature vector $\mathbf{x}$. A two-class classifier C_1 distinguished between coronary calcifications representing the positive class $y_{pos} = y_{LAD} \cup y_{LCX} \cup y_{RCA}$ and all other candidates making up the negative class (y_{neg}). A multiclass classifier C_2 distinguished between all four classes ($y_{LAD}, y_{LCX}, y_{RCA}, y_{neg}$). Both the two-class and

the multiclass classifier consisted of an ensemble of T randomized decision trees (Extra-Trees) [47]. During *training*, each decision tree was trained independently by recursive splitting of a sample set. At each node, the feature-cutoff pair that maximized the normalized information gain was selected from K randomly drawn pairs and the sample set was split accordingly. A leaf node was created when there were n_{min} samples or only one class left in the set. The class distribution of the remaining samples determined a posterior probability distribution $p(Y|\mathbf{x})$ for that node. During *testing*, a candidate's feature vector $\mathbf{x}$ was propagated down each decision tree, ending up in T leaf nodes. The posterior probability distributions corresponding to these nodes were averaged, resulting in a posterior probability distribution for the candidate.

A class label was assigned to a candidate based on the maximum value in $p(Y|\mathbf{x})$. To optimize performance of both classifiers with respect to classification accuracy in terms of volume of candidates, operating points were chosen on the two-class and multiclass ROC curves [48]. This resulted in a set of weight coefficients w_y, $y \in Y$, which were applied to $p(Y|\mathbf{x})$ to obtain $p'(y|\mathbf{x}) = w_y p(y|\mathbf{x}), \forall y \in Y$. Probabilities were re-normalized so that $\sum_{y \in Y} p'(y|\mathbf{x}) = 1$, and the class $\hat{y}$ of candidate $\mathbf{x}$ was found by the optimal decision rule,

$$\hat{y} = \arg\max_y p'(y|\mathbf{x}). \tag{2.1}$$

2.3.5 Ambiguity detection

The purpose of ambiguity detection was to identify candidates to which the classifier could not assign a label with high certainty. Such candidates could subsequently be presented to an expert for review.

Based on uncertainty sampling in active learning, the entropy in the weighted posterior probability distribution $p'(Y|\mathbf{x})$ was used as a measure of labeling certainty [49]. The entropy of the normalized distribution $p'(Y|\mathbf{x})$ was computed as

$$H(Y|\mathbf{x}) = -\sum_{y \in Y} p'(y|\mathbf{x}) \log_2 p'(y|\mathbf{x}).$$

A threshold θ_H on $H(Y|\mathbf{x})$ could be set to only select candidates with $H(Y|\mathbf{x}) \geq \theta_H$ for expert review. The remaining candidates with $H(Y|\mathbf{x}) < \theta_H$ would be labeled without review, according to Eq. 2.1. Two ambiguity detectors were developed to guide review of classifier results: detector R_1 for two-class classifier C_1 and detector R_2 for multiclass classifier C_2. Review was performed in a guided setting, i.e. selected candidates were presented to an expert for possible relabeling of assigned labels. Detector R_1 allowed relabeling to $y_{pos} = y_{LAD} \cup y_{LCX} \cup y_{RCA}$ or y_{neg}, detector R_2 to $y_{LAD}, y_{LCX}, y_{RCA}$ or y_{neg}.

Table 2.1: Distribution of candidate calcifications over the training set (237 scans), validation set (136 scans), test set (530 scans), a subset of the test set (156 scans) used to assess second observer performance, and the orCaScore challenge set (40 scans). Candidates belonged either to CAC in the LAD, LCX or RCA, or to the class of negatives.

	LAD	LCX	RCA	Negatives
Training	436	200	289	305,893
Validation	158	64	86	256,767
Test	761	352	545	723,996
Second observer	245	124	195	223,357
orCaScore	117	64	109	51,244

2.3.6 Evaluation

Total CAC was quantified for each patient, based on the automatic labeling of candidates obtained by classifiers C_1 and C_2, and on the labeling obtained after review with ambiguity detectors R_1 and R_2. For multiclass classifier C_2 and ambiguity detector R_2, CAC was also quantified per coronary artery. First, the performance of CAC detection was evaluated in terms of the number and volume (in mm^3) of calcifications [32], and expressed using sensitivity and the average false positive error rate per image. Second, the agreement between reference and automatically determined total patient CAC volume and Agatston score [9], as well as per artery CAC volume, was computed. This agreement was determined using the two-way intraclass correlation coefficient (ICC) for absolute agreement, with 95% confidence interval, and the mean difference and limits of agreement (±1.96 SD) of Bland-Altman analysis. Third, for each patient, a CVD risk category was determined based on the Agatston score (I: 0, II: 1-10, III: 11-100, IV: 101-400, V: >400) [10]. The agreement between reference and automatic CVD risk categorization was assessed using accuracy and Cohen's linearly weighted κ. Finally, second observer annotations were compared with the reference standard for a subset of images, using the aforementioned evaluation criteria.

2.4 Experiments and results

From the set of 914 included scans, 11 scans were selected as atlases: ten for the estimation of coronary artery centerlines and one for the computation of location features. These atlases did not contain any CAC, high noise levels, motion artifacts or anatomical abnormalities. The remaining 903 scans were consecutively split into three sets: (1) a training set of 237 images used to train the classifiers, (2) a validation set of 136 images used to optimize the parameters of the classifiers and ambiguity detectors, and (3) a test set of 530 images used to evaluate the performance of the classifiers and ambiguity detectors. Furthermore, 156 images from the test set were

consecutively selected to compare the performance of the classifiers and ambiguity detectors with that of a second observer. Finally, the method was evaluated with 40 scans from the publicly available challenge on (semi-)automatic coronary calcium scoring (orCaScore)[1]. This enabled a comparison with other methods participating in this challenge as well as an evaluation of the method's performance with images acquired using scanners of four major CT scanner vendors. Table 2.1 lists the number of candidates per class in each of these sets.

2.4.1 Classifier training and parameter selection

Two-class classifier C_1 and multiclass classifier C_2 were trained on the training set. The classifier parameters were chosen in accordance with the recommendations made by Geurts et al. [47]. The smoothing parameter n_{min} was set to 2 (fully grown trees) and $K = \lfloor \sqrt{f} \rfloor = 8$ features were randomly selected at each split node, where f is the total number of features in $\mathbf{x}$. Experiments showed a performance increase in the validation set with the number of trees T, which was consequently set to 512. For both classifiers, ROC analysis found operating points with relatively high weights for the positive classes, and low weights for the negative class.

Each trained classifier provided a ranking of features by importance. The ten most important features for two-class classifier C_1 were the minimum and average distance to any coronary artery, the minimum, average and maximum distance to the LAD, the distance to the heart-center, the maximum, average and standard deviation of a candidate's intensity, and the 0-th order Gaussian filter at scale 0.3 mm. In contrast, for multiclass classifier C_2 the average distance to the LCX and RCA were among the ten most important features, while the distance to the heart-center and the Gaussian filter were not.

2.4.2 Ambiguity detector parameter selection

For both the two-class (R_1) and the multiclass (R_2) ambiguity detector, an entropy threshold θ_H was chosen to identify candidates with $H(Y|\mathbf{x}) \geq \theta_H$ for review. A low threshold would increase the number of detected and hence potentially corrected errors. On the other hand, a low threshold would also increase the number of selected candidates, and the number of reviewed images. Fig. 2.2 shows the number of reviewed images and corrected errors, and the average number of candidates selected per image, for a range of possible threshold values in the validation set. For both ambiguity detectors, θ_H was chosen such that 60% of errors were detected. For R_1, this point corresponded to $\theta_H = 0.70$ and for R_2 to $\theta_H = 0.95$. Reviewing all candidates with $H(Y|\mathbf{x}) \geq \theta_H$ could be time-consuming in practice, e.g. in very noisy scans. Therefore, a maximum of ten candidates, chosen in order of decreasing entropy, were allowed to be reviewed per CT scan.

[1] http://orcascore.isi.uu.nl/

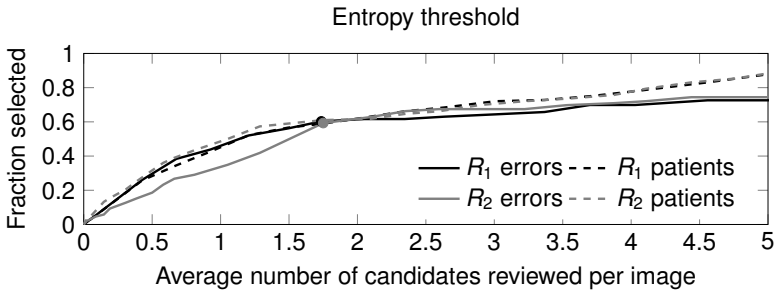


Figure 2.2: The fraction of patients and errors detected by two-class ambiguity detector R_1 and multiclass ambiguity detector R_2, as a function of the average number of candidates selected per image. The origin corresponds to the maximum entropy thresholds θ_H: 1 for R_1 and 2 for R_2. Using these thresholds, no candidates would be selected for review. The dots correspond to the entropy thresholds θ_H used in our experiments: 0.70 for R_1 and 0.95 for R_2.

Figure 2.3: Examples of errors made by two-class classifier C_1. (a) Motion artifact in the RCA detected as a calcification, (b) missed distal calcification in the LCX, (c) missed calcification at origin of the RCA. These errors were all identified by ambiguity detector R_1.

2.4.3 Performance on the test set

The two-class and multiclass classifiers and ambiguity detectors were evaluated using the test set. In this set, review using the ambiguity detector was simulated, i.e. candidates selected for review were relabeled according to the reference standard.

Two-class classification

Two-class classifier C_1 identified 79% of calcifications at the expense of on average 0.2 false positive (FP) errors per image, or 87% of CAC volume with on average 4.1 mm^3 FP volume per image.

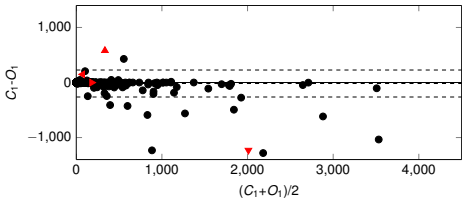
Table 2.2: Agreement between the manually defined reference standard O_1 and the classifiers (two-class C_1, multiclass C_2), ambiguity detectors (two-class R_1, multiclass R_2) and second observer O_2 in (a) the test set (530 scans) and (b) the second observer set (156 scans). For total patient CAC volume and Agatston score, both the two-way intraclass correlation (ICC) for absolute agreement (95% confidence interval) and the mean difference and limits of agreement (±1.96 SD) of Bland-Altman analysis are listed. For CVD risk categorization, accuracy and linearly weighted κ are given.

		O_1-C_1	O_1-C_2	O_1-R_1	O_1-R_2	O_1-O_2
(a)	Volume ICC	0.95 (0.95–0.96)	0.97 (0.97–0.99)	1.00 (1.00–1.00)	0.99 (0.99–1.00)	–
	Volume Bland-Altman	-17 (-227–193)	-4 (-174–165)	-4 (-73–65)	-1 (-83–80)	–
	Agatston ICC	0.96 (0.95–0.96)	0.98 (0.97–0.98)	1.00 (1.00–1.00)	1.00 (0.99–1.00)	–
	Agatston Bland-Altman	-18 (-263–227)	-4 (-198–189)	-4 (-79–71)	-1 (-92–90)	–
	Risk (accuracy / κ)	0.93 / 0.94	0.91 / 0.92	0.99 / 0.99	0.98 / 0.98	–
(b)	Volume ICC	0.93 (0.91–0.95)	0.97 (0.96–0.98)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.99-1.00)
	Volume Bland-Altman	-29 (-303–245)	-15 (-195–166)	-10 (-150–130)	-8 (-134–118)	0 (-97–97)
	Agatston ICC	0.93 (0.90–0.95)	0.97 (0.96–0.98)	0.98 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.99–1.00)
	Agatston Bland-Altman	-31 (-364–301)	-15 (-231–200)	-11 (-181–159)	-8 (-161–144)	1 (-115–117)
	Risk (accuracy / κ)	0.97 / 0.98	0.93 / 0.95	0.97 / 0.98	0.97 / 0.98	0.98 / 0.99

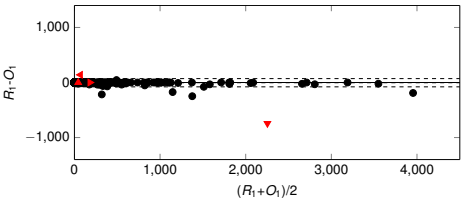
Fig. 2.3 illustrates typical FP and false negative (FN) errors made by the classifier. Common FP errors were aortic wall calcifications close to the origin of the coronary arteries, mitral valve calcifications, high-intensity noise in soft tissues surrounding the coronary arteries, and artifacts caused by artery motion (Fig. 2.3a). The classifier made over 40% of its FP errors in only two scans that contained high noise levels. FN errors most frequently occurred in the distal parts of the arteries (Fig. 2.3b), but also at their origin (Fig. 2.3c).

Table 2.2a lists the agreement between the reference standard (O_1) and the classifier (C_1) for total patient CAC volume, Agatston score and CVD risk categorization. Fig. 2.4a shows a Bland-Altman plot for the agreement between O_1 and C_1 for total patient Agatston score. The confusion matrix presented in Table 2.3a details CVD risk categorization. The classifier assigned 494 (93%) patients to the correct risk category ($\kappa = 0.94$). Miscategorization was often caused by a single misclassified candidate calcification. An overestimation by three CVD risk categories occurred in one image due to calcified pleural plaque closely resembling LAD calcium, and in a second image due to large quantities of high-intensity noise. Conversely, an underestimation by two CVD risk categories occurred in one image due to a missed distal calcification in the LCX (Fig. 2.3b), and in a second image due to a missed calcification at the boundary of the RCA and the ascending aorta (Fig. 2.3c).

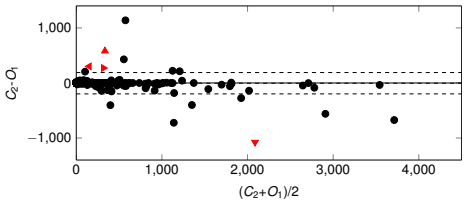
Two-class ambiguity detector R_1 excluded 262/530 images (49%) from review, i.e. these images did not contain any candidates with $H(Y|x) \geq \theta_H$. The excluded images contained only 1% of all classification errors, and 261 out of 262 (99%) images excluded from review were assigned to the correct CVD risk category. In the remaining 51% of images, on average 3.8 candidates were selected for review, including 57% of the FP and FN errors, corresponding to 73% of the misclassified



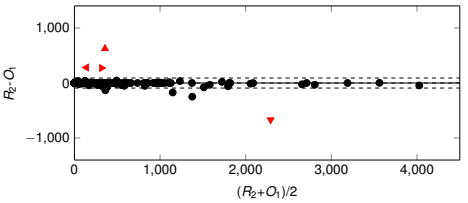
(a) Patient Agatston score O_1 vs. C_1.



(b) Patient Agatston score O_1 vs. R_1.



(c) Patient Agatston score O_1 vs. C_2.



(d) Patient Agatston score O_1 vs. R_2.

Figure 2.4: Bland-Altman plots showing the agreement on total patient Agatston score between the manually defined reference standard (O_1) and (a) two-class classification (C_1) and (b) ambiguity detection with review (R_1), as well as (c) multiclass classification (C_2) and (d) ambiguity detection with review (R_2). Dashed lines represent the mean and the limits of agreement (±1.96 SD) of the differences (Table 2.2a). Four patients for which the difference with the reference score was high after ambiguity detection and review are marked with red triangles in all plots: one patient with an LCX calcification connected to a noise artifact (down-pointing triangle), one patient whose CT contained large amounts of noise (left-pointing triangle), and two patients with aortic wall calcifications close to the origin of the RCA (right-pointing and up-pointing triangles).

volume. Simulated review increased lesion-based sensitivity in the whole test set from 79% to 92% and volume-based sensitivity from 87% to 97%, while the average FP volume per image decreased from 4.1 mm^3 to 1.1 mm^3. Table 2.2a lists the agreement between the reference standard (O_1) and the labels after two-class review (R_1) for total patient CAC volume, Agatston score and CVD risk categorization. Fig. 2.4b shows a Bland-Altman plot for the agreement between O_1 and R_1 for total patient Agatston score. Table 2.3b shows CVD risk categorization according to the labels after review. Fig. 2.5 shows entropy values for several example candidate calcifications.

Multiclass classification

Multiclass classifier C_2 identified 83% of calcifications at the expense of on average 0.4 FP errors per image, or 93% of CAC volume with on average 9.2 mm^3 FP volume per image. Table 2.2a lists the agreement between the reference standard (O_1) and the classifier (C_2) for total patient CAC volume, Agatston score and CVD risk

Table 2.3: Agreement in CVD risk categorization between the reference standard O_1 and (a) two-class classifier C_1 ($\kappa = 0.93$), and (b) review guided by two-class ambiguity detector R_1 ($\kappa = 0.99$).

(a)

Automatic (C_1) \ Reference (O_1)	I	II	III	IV	V	Total
I	**243**	11	1	0	0	255
II	6	**44**	0	1	0	51
III	2	2	**86**	1	0	91
IV	2	0	0	**66**	7	75
V	0	0	1	2	**55**	58
Total	253	57	88	70	62	530

(b)

Review (R_1) \ Reference (O_1)	I	II	III	IV	V	Total
I	**251**	2	0	0	0	253
II	0	**55**	0	0	0	55
III	1	0	**88**	1	0	90
IV	1	0	0	**69**	2	72
V	0	0	0	0	**60**	60
Total	253	57	88	70	62	530

categorization. The classifier assigned 483 (91%) patients to the correct CVD risk category ($\kappa = 0.92$). Miscategorization was caused by errors similar to those generated by the two-class classifier. The multiclass classifier performed slightly below the level of the two-class classifier. Fig. 2.4c shows a Bland-Altman plot for the agreement between O_1 and C_2 for total patient Agatston score.

Table 2.4a shows a confusion matrix for multiclass classification of all candidates in the test set. The majority (90%) of classification errors were FP and FN errors, i.e. misclassification between CAC and negatives, and only 10% of the classification errors were among coronary arteries. Misclassification among coronary arteries occurred exclusively between the LAD and LCX, e.g. at the left coronary bifurcation. Table 2.5a lists the agreement between the classifier and the reference standard on per artery CAC volume.

Multiclass ambiguity detector R_2 excluded 262/530 (49%) images from review. These images contained only 2% of all errors made by the classifier and no errors made among coronary arteries. Moreover, 261/262 (99%) of these images were assigned to the correct CVD risk category. In the remaining 51% of scans, on average 3.6 candidates were selected for review, including 77/190 (41%) FP errors, 143/293 (49%) FN errors and 52/56 (93%) LAD/LCX confusion errors. Table 2.4b shows a

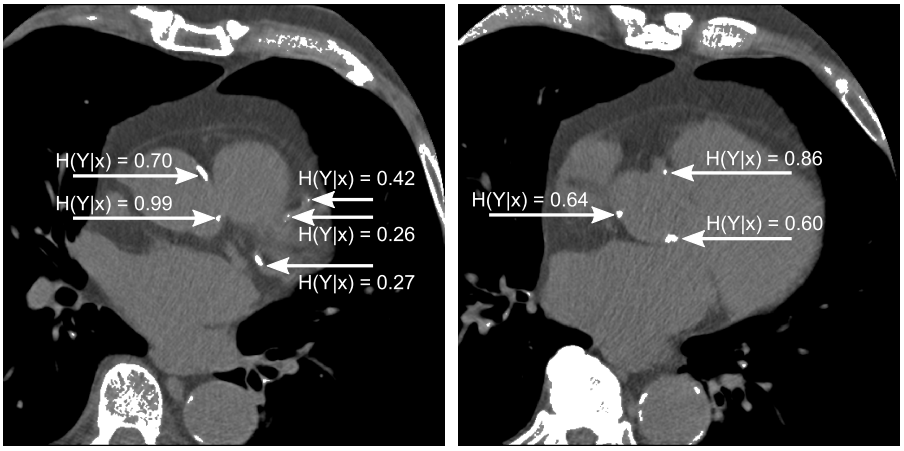


Figure 2.5: Two slices of a cardiac CT image with classification entropy for several candidate calcifications, obtained using two-class classifier C_1. (a) Coronary calcifications with low entropy (0.26–0.42), and a calcification in the left coronary sinus with very high entropy (0.99), (b) a calcification close to the origin of the RCA with high entropy (0.86). Both (a) and (b) show calcifications with elevated entropy (0.60–0.70) in the wall of the ascending aorta. The classification entropy for other negative candidates in these examples was negligible.

Table 2.4: Agreement in candidate calcification labeling between the reference standard O_1 and (a) multiclass classifier C_2, and (b) review guided by multiclass ambiguity detector R_2.

(a)

Automatic (C_2) \ Reference (O_1)	Neg	LAD	LCX	RCA	Total
Neg	**722,148**	54	113	126	722,441
LAD	88	**695**	44	0	827
LCX	10	12	**195**	0	217
RCA	92	0	0	**419**	511
Total	722,338	761	352	545	723,996

(b)

Review (R_2) \ Reference (O_1)	Neg	LAD	LCX	RCA	Total
Neg	**722,225**	14	48	88	722,375
LAD	55	**747**	4	0	806
LCX	4	0	**300**	0	304
RCA	54	0	0	**457**	511
Total	722,338	761	352	545	723,996

Table 2.5: Agreement between the manually defined reference standard (O_1) and the multiclass classifier (C_2) and ambiguity detector (R_2) for per artery CAC volume in (a) the test set (530 scans) and (b) the second observer set (156 scans). The two-way intraclass correlation coefficient (ICC) for absolute agreement (95% confidence interval) is listed as well as the mean difference and limits of agreement (±1.96 SD) of Bland-Altman analysis.

			LAD	LCX	RCA
(a)	O_1-C_2	ICC	0.98 (0.97–0.98)	0.69 (0.64–0.74)	0.95 (0.94–0.96)
		Bland-Altman	3 (-68–73)	-8 (-112–95)	2 (-119–121)
	O_1-R_2	ICC	1.00 (0.99–1.00)	0.95 (0.94–0.96)	0.99 (0.99–0.99)
		Bland-Altman	0 (-33–33)	-2 (-52–48)	1 (-56–57)
(b)	O_1-C_2	ICC	0.97 (0.96–0.98)	0.59 (0.48–0.69)	0.99 (0.98–0.99)
		Bland-Altman	3 (-76–81)	-17 (-190–156)	-1 (-52–51)
	O_1-R_2	ICC	0.98 (0.98–0.99)	0.68 (0.58–0.75)	1.00 (0.99–1.00)
		Bland-Altman	4 (-59–67)	-12 (-170–146)	1 (-34–36)

confusion matrix for multiclass classification of all candidates in the test set after multiclass review. The agreement between the reference standard (O_1) and the labels after multiclass review (R_2) is listed in Table 2.2a for total patient CAC volume, Agatston score and CVD risk categorization, and in Table 2.5a for per artery CAC volume. Fig. 2.4d shows a Bland-Altman plot for the agreement between O_1 and R_2 for total patient Agatston score.

2.4.4 Performance on the second observer set

This set consisted of 156 consecutively selected images from the test set. The performance of a second observer on this set was compared with that of the classifiers and ambiguity detection followed by review. In contrast to the simulated review presented in Section 2.4.3, this review was performed by a human expert.

Two-class classification

Table 2.2b lists the agreement between the reference standard (O_1) and the second observer (O_2), between the reference standard (O_1) and the two-class classifier (C_1), and between the reference standard (O_1) and the labels after two-class review (R_1), for total patient CAC volume, Agatston score and CVD risk categorization. Two-class ambiguity detector R_1 excluded 64/156 images (41%) from review. In the remaining images, on average 4.0 candidates were reviewed. The median time required for two-class expert review was 41 s. Manual scoring was performed per coronary artery, hence the time required for two-class manual scoring is not available.

Table 2.6: Agreement between the manually defined reference standard O_1 and the classifiers (two-class C_1, multiclass C_2) and ambiguity detectors (two-class R_1, multiclass R_2) in scans obtained with GE, Philips, Siemens and Toshiba scanners. For total patient CAC volume and Agatston score, both the two-way intraclass correlation (ICC) for absolute agreement (95% confidence interval) and the mean difference and limits of agreement (±1.96 SD) of Bland-Altman analysis are listed. For CVD risk categorization, accuracy and linearly weighted κ are listed.

		O_1-C_1	O_1-C_2	O_1-R_1	O_1-R_2
GE	Volume ICC	0.96 (0.87–0.99)	0.96 (0.87–0.99)	1.00 (0.99–1.00)	0.99 (0.96–1.00)
	Volume Bland-Altman	-25 (-226–176)	-19 (-227–189)	-11 (-73–52)	-20 (-137–98)
	Agatston ICC	0.97 (0.90–0.99)	0.97 (0.90–0.99)	1.00 (1.00–1.00)	0.99 (0.98–1.00)
	Agatston Bland-Altman	-23 (-220–174)	-16 (-222–190)	-7 (-50–36)	-16 (-113–81)
	Risk (accuracy / κ)	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
Philips	Volume ICC	1.00 (0.99–1.00)	0.99 (0.98–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Volume Bland-Altman	2 (-19–23)	-16 (-115–83)	-1 (-6–4)	-3 (-19–13)
	Agatston ICC	1.00 (1.00–1.00)	1.00 (0.98–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Agatston Bland-Altman	3 (-22–28)	-16 (-117–85)	-1 (-4–3)	-2 (-11–8)
	Risk (accuracy / κ)	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
Siemens	Volume ICC	1.00 (0.99–1.00)	1.00 (0.98–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Volume Bland-Altman	-6 (-28–17)	-2 (-36–32)	-3 (-13–7)	-2 (-8–4)
	Agatston ICC	1.00 (0.99–1.00)	1.00 (0.99–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Agatston Bland-Altman	-4 (-30–21)	-2 (-33–29)	-1 (-16–4)	-1 (-5–3)
	Risk (accuracy / κ)	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
Toshiba	Volume ICC	0.98 (0.92–0.99)	0.98 (0.91–0.99)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Volume Bland-Altman	17 (-43–76)	15 (-54–84)	-1 (-3–1)	-1 (-3–1)
	Agatston ICC	0.98 (0.92–1.00)	0.98 (0.91–0.99)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Agatston Bland-Altman	20 (-48–88)	20 (-62–101)	0 (-1–0)	0 (-1–0)
	Risk (accuracy / κ)	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00

Multiclass classification

The ICC between the reference standard (O_1) and the second observer (O_2) for per artery CAC volume was 0.99 (0.99–1.00), 0.91 (0.87–0.93) and 1.00 (0.99–1.00) for LAD, LCX and RCA, respectively. Table 2.2b lists the agreement between the reference standard (O_1) and the classifier (C_2), as well as the agreement between the reference standard (O_1) and labels after multiclass review (R_2) for total patient CAC volume, Agatston score and CVD risk categorization. Table 2.5b lists the agreement between the reference standard (O_1) and the multiclass classifier (C_2), as well as the agreement between the reference standard (O_1) and the labels after multiclass review (R_2) for per artery CAC volume. Multiclass ambiguity detector R_2 excluded 64/156 images (41%) from review. In the remaining images, on average 3.6 candidates were reviewed. Multiclass expert review was substantially faster than manual scoring, median time 45 s vs. 128 s. per image.

Table 2.7: Agreement between the manually defined reference standard (O_1) and the multiclass classifier (C_2) and ambiguity detector (R_2) for per artery CAC volume in GE, Philips, Siemens and Toshiba scans. The two-way intraclass correlation coefficient (ICC) for absolute agreement (95% confidence interval) is listed as well as the mean difference and limits of agreement (±1.96 SD) of Bland-Altman analysis.

		O_1-C_2			O_1-R_2		
		LAD	LCX	RCA	LAD	LCX	RCA
GE	ICC	0.96 (0.87–0.99)	0.99 (0.97–1.00)	0.95 (0.81–0.99)	1.00 (1.00–1.00)	0.99 (0.96–1.00)	0.96 (0.84–0.99)
	Bland-Altman	-11 (-115–93)	-3 (-30–23)	-5 (-91–82)	-2 (-16–12)	-6 (-38–27)	-11 (-82–59)
Philips	ICC	1.00 (1.00–1.00)	1.00 (0.99–1.00)	0.98 (0.91–0.99)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Bland-Altman	-1 (-6–5)	-1 (-14–11)	-14 (-103–74)	0 (0–0)	-1 (-10–7)	-1 (-9–6)
Siemens	ICC	0.99 (0.96–1.00)	0.87 (0.58–0.97)	0.98 (0.92–0.99)	1.00 (1.00–1.00)	1.00 (0.99–1.00)	1.00 (1.00–1.00)
	Bland-Altman	2 (-34–37)	-5 (-41–31)	1 (-25–27)	0 (0–0)	-1 (-6–4)	-1 (-6–3)
Toshiba	ICC	0.99 (0.95–1.00)	0.91 (0.66–0.98)	0.87 (0.59–0.97)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Bland-Altman	-8 (-25–40)	0 (-41–41)	8 (-40–55)	0 (-1–1)	0 (0–0)	0 (-3–2)

2.4.5 Performance on CTs acquired with different scanners

To evaluate the performance of the proposed method with images acquired using different CT scanners, data available within the orCaScore challenge was used. This challenge provides a set of CSCT and corresponding CCTA scans of 40 patients, evenly distributed over CVD risk categories. The scans were made using a scanner from one of four major CT vendors (GE, Philips, Siemens, Toshiba). All images were made using a standardized calcium scoring protocol but they were reconstructed using a vendor-specific reconstruction kernel: standard (GE), XCA (Philips), B35f (Siemens), FC12 (Toshiba). A detailed data description can be found on the challenge website. The two-class and multiclass classifiers and ambiguity detectors were evaluated. Review using the ambiguity detector was simulated, as described in Section 2.4.3.

Two-class classification

Table 2.6 lists the agreement between the reference standard (O_1) and the two-class classifier (C_1), and between the reference standard (O_1) and the labels after two-class review (R_1) for total patient CAC volume, Agatston score and CVD risk for each scanner vendor separately. Two-class ambiguity detector R_1 excluded 11/40 (28%) scans from review and selected 4.6 candidates per reviewed image. These included 60% of errors.

Multiclass classification

Table 2.6 lists the agreement between the reference standard (O_1) and the multiclass classifier (C_2), and between the reference standard (O_1) and the labels after multiclass review (R_2) for total patient CAC volume, Agatston score and CVD risk for each scanner vendor separately. Table 2.7 lists the agreement between the reference

Table 2.8: Comparison of methods for automatic coronary calcium scoring, based on previously published results (top) and within the orCaScore challenge framework (bottom). For each study, the type (Test data: calcium scoring CT [CSCT], coronary CT angiography [CCTA] or low-dose non-contrast enhanced chest CT [Chest CT]) and number of scans (# Scans) with which the method was evaluated are listed. The reported CVD risk categorization accuracy per patient (Risk), correlation between reference and automatically computed calcium scores (Corr.) and sensitivity (Sens.), positive predictive value (PPV) and FP rate per scan (FP/scan) for lesion identification are shown. Different correlation metrics were reported: ^{s} Spearman's ρ, ^{p} Pearson's ρ, ^{r} linear regression R-squared and ^{i} the intraclass correlation coefficient for absolute agreement. Results presented in the top part used different image sets, different CVD risk categories and different correlation metrics.

Study	Test data	# Scans	Risk (%)	Corr.	Sens. (%)	PPV (%)	FP/scan
Išgum et al. [33]	CSCT	76	93	–	73.8	–	0.1
Kurkure et al. [34]	CSCT	105	–	–	92.1	–	4.7
Brunner et al. [35]	CSCT	30	–	–	86.3	–	–
Shahzad et al. [13]	CSCT	157	93	0.97^{P}	83.9	–	1.3
Saur et al. [36]	CSCT + CCTA	127	–	–	86.3	87.8	–
Išgum et al. [14]	Chest CT	231	82	0.93^{S}	58.6	–	0.1
Xie et al. [37]	Chest CT	41	80	0.91^{R}	–	–	–
Proposed method	CSCT	530	93	$0.96\ (0.95–0.96)^{I}$	79.0	92.0	0.2
Shahzad et al.	CSCT + CCTA	40	88	$0.97\ (0.95–0.98)^{I}$	62.3	87.9	–
Yang et al.	CSCT + CCTA	40	98	$0.99\ (0.98–1.00)^{I}$	94.0	95.5	–
Kelm et al.	CSCT + CCTA	40	95	$0.98\ (0.96–0.99)^{I}$	83.9	95.3	–
Kondo	CSCT + CCTA	40	80	$0.62\ (0.39–0.78)^{I}$	51.6	64.7	–
Proposed method	CSCT	40	100	$0.99\ (0.97–0.99)^{I}$	84.5	95.0	–

standard (O_1) and the multiclass classifier (C_2), as well as the agreement between the reference standard (O_1) and labels after multiclass review (R_2) for per artery CAC volume. Multiclass ambiguity detector excluded 7/40 (18%) scans from review and selected 4.7 candidates per reviewed image. These included 58% of errors.

2.4.6 Comparison with other methods

The performance of the proposed method was compared with other previously published algorithms and with algorithms participating in the orCaScore challenge. A direct comparison with methods that did not participate in the challenge cannot be made due to differences in the used data and the performed evaluation. Hence, Table 2.8 lists the data used by each method, the size of the evaluated data set and the reported results. For completeness, algorithms performing the scoring in cardiac and chest CT scans are included. Even though the listed results cannot be directly compared, they provide an indication of the differences. Furthermore, Table 2.8 lists results of the methods that participated in the orCaScore challenge. Detailed results and a description of each method are available at the challenge website. These methods were evaluated using the same data and performance criteria. Hence, these

results can be directly compared. However, note that the authors of the proposed method had access to the reference annotations.

2.5 Discussion

A system for automatic coronary calcium scoring has been presented. The system determined both total patient and per artery coronary calcium burden, using only a patient's CSCT. The results demonstrated that the proposed method is able to accurately detect CAC volume (87% sensitivity, on average 4.1 mm^3 FP per image) and assign patients to a CVD risk category ($\kappa = 0.94$). Nevertheless, performance of the proposed automatic method was below that of the second observer ($\kappa = 0.99$). This might prohibit the use of fully automatic calcium scoring in clinical practice. Therefore, ambiguity detectors were developed to guide review of candidate calcifications which the classifiers could not label with high certainty. Review of selected candidates in a limited set of scans (51%) increased agreement with the reference standard to the level of the second observer, with little expert effort. This allows both accurate and fast calcium scoring.

Discrimination between patients with zero and non-zero CAC scores is of high clinical relevance as zero calcium score could serve as a gatekeeper for invasive diagnostic procedures [50] or prescription of drugs lowering the risk of a CVD event [51, 52]. The proposed automatic system incorrectly assigned a zero CAC score instead of a positive score and vice versa in 22 out of 34 miscategorized patients. The twelve patients to whom a zero CAC score was incorrectly assigned by our method had low reference scores, ranging from 0.8 to 23.8, median (interquartile range) 1.8 (1.1–5.5). Conversely, in the ten patients incorrect assignment of a positive calcium score was caused by calcified pleural plaque (1), mitral valve calcification (1), ascending aortic wall calcification (2), motion artifacts in the RCA (2), or noise artifacts (4). Ambiguity detection and subsequent guided review would ensure that most of these patients (18/22) would be reassigned to their correct CVD risk category.

Ambiguity detection was based on the classification entropy, which requires accurate estimates of the true posterior probabilities. Forests of randomized decision trees have been shown to provide such estimates [53]. Our experiments showed that based solely on a single entropy threshold, the majority (57%) of misclassified calcifications were detected. In addition, in our experiments, a maximum was placed on the number of candidates that could be selected for review in each image, selecting only the most ambiguous candidates. While this prevented laborious relabeling of a large number of ambiguous candidates in scans with high noise levels, it could also lead to overestimated CAC scores caused by uncorrected FP errors in such scans. In fact, a high number of candidates eligible for review in an image could be an indication for fully manual scoring. Future work could investigate the range of entropy thresholds with which review might be beneficial in terms of performance increase and workload reduction. Furthermore, the criteria used by the ambiguity detec-

tors could be optimized for specific purposes such as the management of cardiac patients, where the aim might be to achieve very accurate scoring, or screening for CVD, where an emphasis might be placed on CVD risk categorization.

The primary clinical purpose of calcium scoring is the quantification of total calcified coronary plaque in an individual patient. For this, the two-class classifier, which identified CAC irrespectively of its coronary artery, showed better performance than the multiclass classifier, which labeled CAC according to the artery. Obviously, the determination of per artery CAC burden involves more classes, making it a different and likely more complex task. This was demonstrated by the features that were ranked highest by each classifier. The most important features for the two-class classifier were those describing the distance to the coronary artery tree, that is, to any of the coronary arteries, while the multiclass classifier placed more emphasis on the distance to each specific coronary artery. Furthermore, the supervised nature of the proposed method requires a substantial number of representative training samples. Given the fixed size of the available training set, the multiclass classifier had fewer training samples for each class. Enrichment of the training data with additional samples might eliminate the difference between two-class and multiclass classifier performances. In addition, enrichment of the training data might allow further separation of per-artery class labels, like those of the LM trunk and LAD, which were in the current study considered as one class. Likewise, the RCA class label could be separated to independently label calcifications located in the proximal part and in the crux. Due to varying anatomy in these regions, it is likely that a very large training set would need to be available. Continuous integration of new labeled samples through an active learning approach would enable the inclusion of rarely appearing calcifications in the training data, and therefore, might be an interesting extension to the guided review method proposed here.

The ranking of features in the two-class and multiclass classifiers emphasized the importance of location features, which was demonstrated in previous work [14]. To accurately compute location features, the method proposed in [36] and several methods evaluated in the orCaScore challenge used CCTA scans corresponding to CSCT scans (Table 2.8). In practice, an additional CCTA image might not always be available, and hence it might be advantageous to use only CSCT. However, when using only CSCT, the position of coronary arteries can at best be estimated. Visual inspection of the obtained estimates revealed their occasional suboptimal estimation, mostly in patients with rarely appearing anatomical variations or abnormalities. Although the presented results demonstrated that the quality of these estimations was sufficient for the accurate identification of CAC, further improvement in estimation of the location of the coronary arteries might most likely lead to a further increase in performance. The manual annotation of coronary arteries in the CCTA atlas images as it is performed in this study is accurate but time-consuming. Hence, the number of used atlases was limited. Future work could investigate the application of automatic segmentation of coronary arteries in CCTA using one of the available

methods [54, 55]. Additionally, this would allow the utilization of a larger number of CCTA atlases representing a range of anatomical variations (e.g. left and right dominance, commonly appearing branches). This might lead to better estimation of the coronary artery location in CSCT.

In addition to the test set, the algorithm was evaluated using images from the orCaScore challenge, acquired on CT scanners from four different vendors. It has been shown that images made with CT scanners from different vendors yield significantly different calcium scores [56]. Nevertheless, the results indicated that the impact of these differences on the proposed method was small; automatic CVD risk categorization was excellent in images from all CT scanners. While some differences between per-vendor results were visible, the number of scans per vendor (10) was too small to determine whether these differences were caused by the method, patient characteristics or scanner vendor. Furthermore, the results obtained with the images from the orCaScore challenge compared favorably with those obtained with the test set (per lesion sensitivity 85% vs. 79%, positive predictive value 95% vs. 92%). This might be a consequence of the image selection criteria. Namely, images in the orCaScore challenge were carefully selected to exclude scans with motion or noise artifacts, while such scans were included in the test set of this study. Moreover, the images in the challenge were evenly distributed over CVD risk categories. Therefore, they contained more patients with non-zero CAC scores (80% vs. 52%) and on average more calcifications per image (7.3 vs. 3.1) than scans in the test set of this study. Hence, in the orCaScore set, two-class ambiguity detector R_1 selected more scans for review (72% vs. 51%) and selected more candidates per reviewed images (4.6 vs. 3.8) than in the test set of this study. Nevertheless, the percentage of errors selected for review was similar in both sets (60% vs. 57%).

2.6 Conclusion

An algorithm for the automatic identification and quantification of CAC in routinely acquired non-contrast-enhanced ECG-triggered cardiac CT was presented. The system was able to accurately determine total patient and per artery CAC burden. Detection of ambiguity in automatic labeling and subsequent guided review of selected candidate calcifications in 51% of patients increased the system's performance to that of the second observer with little expert effort. This might enable the application of automatic CAC scoring in clinical settings.

Chapter 3

An evaluation of automatic coronary artery calcium scoring methods with cardiac CT using the orCaScore framework

Based on:
J.M. Wolterink, T. Leiner, B.D. de Vos, J.L. Coatrieux, B.M. Kelm, S. Kondo, R.A. Salgado, R. Shahzad, H. Shu, M. Snoeren, R.A.P. Takx, L.J. van Vliet, T. van Walsum, T.P. Willems, G. Yang, Y. Zheng, M.A. Viergever, I. Išgum. "An Evaluation of Automatic Coronary Artery Calcium Scoring Methods with Cardiac CT using the orCaScore Framework," *Medical Physics*, 2016, vol. 43, pp. 2361–2373

Abstract

The amount of coronary artery calcification (CAC) is a strong and independent predictor of cardiovascular disease (CVD) events. In clinical practice, CAC is manually identified and automatically quantified in cardiac CT using commercially available software. This is a tedious and time-consuming process in large-scale studies. Therefore, a number of automatic methods that require no interaction and semi-automatic methods that require very limited interaction for the identification of CAC in cardiac CT have been proposed. Thus far, a comparison of their performance has been lacking. The objective of this study was to perform an independent evaluation of (semi-)automatic methods for CAC scoring in cardiac CT using a publicly available standardized framework.

Cardiac CT exams of 72 patients distributed over four CVD risk categories were provided for (semi-)automatic CAC scoring. Each exam consisted of a non-contrast-enhanced calcium scoring CT (CSCT) and a corresponding coronary CT angiography (CCTA) scan. The exams were acquired in four different hospitals using state-of-the-art equipment from four major CT scanner vendors. The data were divided into 32 training exams and 40 test exams. A reference standard for CAC in CSCT was defined by consensus of two experts following a clinical protocol. The framework organizers evaluated the performance of (semi-)automatic methods on test CSCT scans, per lesion, artery and patient.

Five (semi-)automatic methods were evaluated. Four methods used both CSCT and CCTA to identify CAC, one method used only CSCT. The evaluated methods correctly detected between 52% and 94% of CAC lesions with positive predictive values between 65% and 96%. Lesions in distal coronary arteries were most commonly missed and aortic calcifications close to the coronary ostia were the most common false positive errors. The majority (between 88% and 98%) of correctly identified CAC lesions were assigned to the correct artery. Linearly weighted Cohen's kappa for patient CVD risk categorization by the evaluated methods ranged from 0.80 to 1.00.

A publicly available standardized framework for the evaluation of (semi-)automatic methods for CAC identification in cardiac CT is described. An evaluation of five (semi-)automatic methods within this framework shows that automatic per patient CVD risk categorization is feasible. CAC lesions at ambiguous locations such as the coronary ostia remain challenging, but their detection had limited impact on CVD risk determination.

3.1 Introduction

Cardiovascular disease (CVD) is the global leading cause of death [1]. The amount of coronary artery calcification (CAC) is a strong and independent predictor of CVD events [7, 8, 57, 10, 58]. Therefore, in clinical practice CAC is routinely identified and quantified in CT of the heart, most commonly in non-contrast-enhanced ECG-triggered cardiac calcium scoring CT (CSCT). In calcium scoring, a human operator manually identifies CAC lesions in each image slice using commercially available software, which subsequently quantifies the manually identified lesions. Manual identification of CAC is not considered a difficult task, but it is time-consuming and impractical in large-scale (epidemiological) studies. Moreover, the number of cardiac CT exams is expected to rise as recent guidelines recommend CT-based CAC scoring in adults at intermediate and low-to-intermediate risk of a CVD event [12]. As an alternative to manual CAC lesion identification, software that requires no or very limited interaction, i.e. (semi-)automatic calcium scoring methods, would reduce the workload of experts and enable large-scale studies.

In the research literature, a number of (semi-)automatic methods for coronary calcium scoring in CT have been proposed (Table 3.1). In order to assess the clinical potential of these methods, a detailed comparison of their performance is required. However, a comparison based purely on published results does not provide a reliable assessment for several reasons. First, methods have been evaluated using different data sets, in terms of image acquisition, image reconstruction and patient inclusion. CAC scoring methods have been developed for diverse CT images of the heart, namely CSCT [13, 15, 59, 35, 34, 33, 60, 61], contrast-enhanced coronary CT angiography (CCTA) [62, 63, 64, 65, 66, 67], a combination of CSCT and CCTA [36] or non-contrast-enhanced non-ECG-triggered chest CT [14, 37]. These scans may all be used to determine the amount of CAC, but their characteristics pose different challenges. CCTA provides excellent visibility of the coronary arteries, but it may be hard to differentiate between CAC and coronary lumen. On the other hand, while the contrast between CAC and coronary lumen is high in non-contrast-enhanced CT, it is practically impossible to localize the coronary arteries, particularly in non-ECG-triggered chest CT. In addition, images have been acquired on scanners from different vendors, using either electron beam CT (EBCT) or multi-detector CT (MDCT), which may yield different image characteristics. The number of included patients per study ranged from 10 to 530, including lung cancer screening participants [14] and patients with known or suspected CVD [63]. Hence, anatomical characteristics and CAC distribution might differ across studies. Second, the CAC scoring task was described and evaluated differently across studies. While most methods identified CAC irrespective of its location in the coronary artery, several studies also performed the more challenging task of CAC labeling per artery [13, 15, 60, 35]. Studies have used slightly different definitions of CAC lesions, including variations in HU threshold and limitations on the required minimum lesion size. Moreover, a range of quantitative evaluation results have been reported, including different correlation

Table 3.1: Published evaluations of methods for automatic coronary calcium scoring, with most recent publications listed first. Reported evaluation: scans required for analysis (Scans; non-contrast calcium scoring CT [CSCT], coronary CT angiography [CCTA], chest CT), CT scanner type (CT type; multi detector CT [MDCT], electron beam CT [EBCT]) and vendor (CT vendor; GE, Philips, Siemens, Toshiba), number of test patients (Patients), and reported evaluation (Evaluation; per lesion, per artery, per patient).

	Scans	CT type	CT vendor	Patients	Evaluation
Schuhbaeck et al. [68][1]	CCTA	MDCT	Siemens	44	Patient
Wolterink et al. [62]	CCTA	MDCT	Philips	50	Patient
Wolterink et al. [15]	CSCT	MDCT	Philips	530	Lesion, artery, patient
Ding et al. [60]	CSCT	EBCT, MDCT	GE, Siemens	50	Artery, patient
Ahmed et al. [63]	CCTA	MDCT	Toshiba	100	Patient
Xie et al. [37]	Chest CT	–	–	41	Patient
Eilot et al. [64]	CCTA	MDCT	Philips, Siemens	263	Lesion, patient
Takx et al. [69][2]	Chest CT	MDCT	Philips	1749	Patient
Shahzad et al. [13]	CSCT	MDCT	Siemens	157	Lesion, artery, patient
Wu et al. [61]	CSCT	–	–	16	Patient
Arnold et al. [59]	CSCT	MDCT	GE	78	Patient
Mittal et al. [65]	CCTA	–	–	165	Lesion
Kurkure et al. [34]	CSCT	EBCT	–	105	Lesion
Brunner et al. [35]	CSCT	EBCT	GE	30	Lesion, artery
Saur et al. [36]	CCTA, CSCT	MDCT	Siemens	127	Lesion
Išgum et al. [33]	CSCT	MDCT	Philips	76	Lesion, patient
Wesarg et al. [67]	CCTA	MDCT	Siemens	10	Lesion, patient

coefficients and classification performance measures. Finally, re-implementation of published methods for a comparison using standardized data and evaluation criteria would hardly be feasible due to their complexity.

In this study, we present a publicly available framework for the evaluation of (semi-)automatic methods for c**or**onary artery **ca**lcium **scor**ing (orCaScore) which was set up to overcome these issues. In this framework, corresponding CSCT and CCTA scans of a stratified set of patients are provided. This set consists of both male and female patients distributed over four CVD risk categories, who were scanned on scanners from four major CT vendors, acquired in four different hospitals. A detailed task description and evaluation criteria were provided. Researchers were invited to analyze the provided exams available within this study using their own research or commercially available (semi-)automatic methods. Evaluation was performed by the framework organizers. This work describes the orCaScore framework and an independent comparison of five methods for (semi-)automatic CAC scoring.

[1] Method description provided by Dey et al. [66]
[2] Method description provided by Išgum et al. [14]

3.2 Materials and methods

3.2.1 orCaScore evaluation framework

The orCaScore evaluation framework is web-based[3]. The website provides a detailed description of the task, available data and evaluation protocol. Using the website, researchers from academia and industry can register to evaluate their (semi-)automatic calcium scoring method. After registration, clinically obtained cardiac CT exams with pairs of CSCT and corresponding CCTA scans may be downloaded. These exams have been divided into a training set and a test set. A reference standard is available to participants for training exams but not for test exams. The task is to identify CAC lesions in the test exams and to optionally label each CAC lesion according to the coronary artery it is located in. For this, participants use their own research or commercial software which is either fully automatic or semi-automatic, i.e. requiring very limited manual interaction. Obtained results are then submitted for evaluation, performed by the study organizers using a standardized set of performance criteria.

The framework was launched at a workshop organized in conjunction with the 2014 Medical Image Computing and Computer Assisted Intervention (MICCAI) conference in Boston, MA. At this workshop, an additional set of test exams was provided for on-site analysis. Potential participants were notified of the framework and workshop by personal invitation, via an image analysis e-mail list and by publicity provided by the MICCAI organizers. This work describes an evaluation of four methods that were presented at the workshop and one method developed by the framework organizers. Here, combined results on both the web-based and the workshop test set are presented. The web-based evaluation framework remains open for future submissions of new or updated methods, results of which will be posted on the website.

3.2.2 Data

The framework provides cardiac CT exams of patients with both a CSCT and a corresponding CCTA, from four academic hospitals (Antwerp University Hospital, Antwerp, Belgium; Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands; University Medical Center Groningen, Groningen, The Netherlands; and University Medical Center Utrecht, Utrecht, The Netherlands). Patients were considered for inclusion in consecutive scanning order. Those patients with anatomical abnormalities, intracoronary stents, and metal implants as well as CTs showing severe motion artifacts or extremely high levels of noise determined by visual inspection were excluded. Eighteen (nine male, nine female) patients were included per hospital for a total of 72 exams. Patient inclusion was stratified according to CVD risk categories based on Agatston scores, with cut-off points as proposed by

[3] http://orcascore.isi.uu.nl/

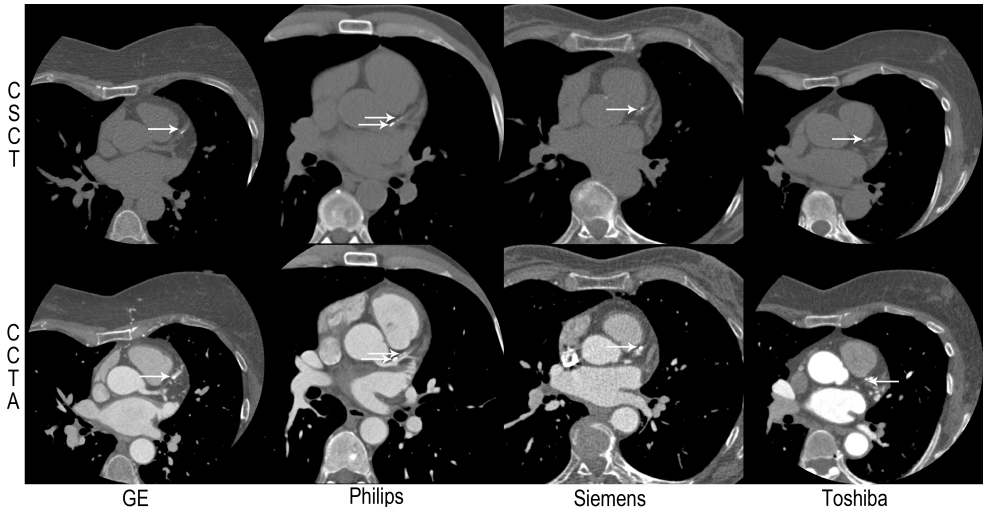


Figure 3.1: CSCT and corresponding CCTA scans of four patients contained in the training set, scanned on a GE, Philips, Siemens and Toshiba scanner, respectively. Each of these patients had CAC in the left coronary artery tree, marked with white arrows. Window and level are the same for all images.

Detrano et al. [7] (I: 0, II: 1-100, III: 101-300, IV: >300). The web-based training and test exams both included images of 32 patients: one male and one female patient in each CVD risk category from each site. The test exams provided at the workshop included images of eight patients: one patient in category II and one in category IV from each site. This set was limited in size to allow on-site analysis during the workshop.

For each patient, a CCTA and a CSCT scan were acquired in one session as part of clinical routine. Patients were scanned on a multi-detector row CT scanner from one of four different vendors (Lightspeed VCT, software version GMP VCT.26, GE Healthcare, Milwaukee, Wisconsin; Brilliance iCT, software version 3.2.0, Philips Healthcare, Best, The Netherlands; SOMATOM Definition Flash, software version Syngo CT 2010A, Siemens Healthcare, Forchheim, Germany; and Aquilion ONE, software version 4.93, Toshiba Medical Systems, Otawara, Japan). All acquisitions were synchronized to the diastolic rest period by ECG-triggering, at 70% (GE, Siemens), 75% (Toshiba) or 78% (Philips) of the R-R interval. Fig. 3.1 shows example CSCT and CCTA images for four patients with CAC in the left coronary artery tree, who were each scanned on a different scanner.

The CSCTs were acquired with a vendor-specific sequential scanning protocol and a tube voltage of 120 kVp. Images with 0.4 x 0.4 - 0.5 x 0.5 mm^2 in-plane resolution and 2.5 mm (GE) or 3 mm (Philips, Siemens, Toshiba) section thickness and increment were reconstructed using vendor-specific software, with the following kernels: standard (GE), XCA (Philips), B35f (Siemens), and FC12 (Toshiba).

The CCTAs were acquired with a vendor-specific sequential scanning protocol, contrast bolus-tracking and a tube voltage of 100 or 120 kVp. Images with 0.4 x 0.4 - 0.5 x 0.5 mm^2 in-plane resolution and 0.625/0.625 mm (GE), 0.9/0.45 mm (Philips), 0.6/0.4 mm (Siemens) and 0.5/0.25 mm (Toshiba) section thickness/increment were reconstructed using vendor-specific software, with the following kernels: standard (GE), XCA (Philips), B30f (Siemens), and FC03 (Toshiba).

Information about the hospitals or scanners from which CTs originated or the distribution of patients with respect to sex or CVD risk category was not provided to study participants.

3.2.3 Reference standard

Each exam contained a CSCT and a CCTA scan. CSCT is the de facto standard for CAC identification and quantification and hence, evaluation was limited to CAC identification in CSCT. Therefore, a reference standard for CAC was set in all provided CSCT scans using annotations by two expert observers: a radiologist (TL) with 12 years of experience in CAC scoring (>5000 CSCT scans) and a research physician (RAPT) with 5 years of experience in CAC scoring (>1000 CSCT scans). Thresholding at 130 HU identified potential CAC lesions (6-connected components with volume >1.5 mm^3). These included not only CAC, but also other high-density lesions such as bony structures and aortic calcifications. Both observers independently identified each CAC lesion and labeled it according to location (left anterior descending artery [LAD] including the left main stem, left circumflex artery [LCX], or right coronary artery [RCA]). In a subsequent joint reading session, the observers resolved labeling differences by consensus. Annotations after consensus served as the reference standard. All annotations were made using custom software (iX Viewer, Image Sciences Institute, Utrecht, The Netherlands) which showed the CSCT scan synchronized with the corresponding CCTA scan for reference.

3.2.4 Evaluated methods

Four methods (A, B, C, D) participated in the workshop. A fifth method (E) was developed by the framework organizers, who had access to all study data and the reference standard. Table 3.2 lists the required CT scan, user interaction, provided labels for identified lesions and the average processing runtime for each method. All methods required both CSCT and CCTA scans for CAC identification, except for method E, which only used CSCT scans. Methods A, C and E were automatic, while the other two methods were semi-automatic, requiring manual interaction in all (method D) or a limited number (method B) of exams. The evaluated methods typically identified CAC in two subsequent stages: in the first stage information about the location of the coronary arteries was obtained, in the second stage this information was used to identify CAC lesions. Here, we provide a brief description

Table 3.2: Methods evaluated in this study: scans required for analysis (Scans; non-contrast calcium scoring CT [CSCT], coronary CT angiography [CCTA]), required user interaction (Interaction), labels provided for individual lesions (Labels; left anterior descending artery [LAD], left circumflex artery [LCX] and right coronary artery [RCA]) and average runtime per exam (Time).

	Scans	Interaction	Labels	Time
Method A	CSCT + CCTA	No interaction required	LAD, LCX, RCA	80 min
Method B	CSCT + CCTA	Optional correction of automatic ostia detection	LAD, LCX, RCA	8 min
Method C	CSCT + CCTA	No interaction required	LAD, LCX, RCA	2 min
Method D	CSCT + CCTA	Manual seed placement in the ascending aorta	LAD, LCX, RCA	3 min
Method E[4]	CSCT	No interaction required	LAD, LCX, RCA	20 min

of each method. A detailed description of each evaluated method is available on the orCaScore website.

Method A - Erasmus Medical Center (Rotterdam, The Netherlands)

CAC was identified by analyzing the test CCTA scan and the test CSCT scan. First, a standardized heart coordinate system and patient-specific probabilistic maps for the occurrence of coronary arteries were obtained for the test CCTA. A probabilistic map contained a probability for each voxel to be part one of the three major coronary arteries (LAD, LCX and RCA). Eight CCTA atlas scans, each with a heart coordinate system and a probabilistic map, were registered to the test CCTA scan using rigid and subsequent non-rigid registration with *elastix* [42]. The coordinate systems and probabilistic maps were transformed from all atlas CCTA scans to the test CCTA scan and combined through averaging. Next, the test CCTA scan was registered to the test CSCT scan. The heart coordinate system and probabilistic map were transformed from the test CCTA scan to the test CSCT scan. This provided an estimate of the location of the coronary arteries in the test CSCT. Finally, CAC was identified using supervised pattern recognition. All candidate CAC lesions (> 130 HU, > 1.5 mm^3) in the CSCT were extracted and described by volume, intensity features, coordinates in the heart coordinate system, and the coronary artery probability from the probabilistic map. Candidates were classified as CAC or non-CAC using a 9-nearest neighbor classifier trained with 76 previously acquired CSCT scans. Artery labels of identified CAC lesions were determined using the patient-specific probabilistic map. This method was based on previously described work [13].

[4]Framework organizer with access to the reference standard

Method B - Southeast University (Nanjing, China)

CAC was identified by analyzing the test CCTA scan and the test CSCT scan. The test CCTA scan was preprocessed by cylindrical cropping around an initial automatic segmentation of the ascending aorta obtained with circular Hough transforms. Next, the heart and the ascending aorta in the test CCTA were segmented using eight CCTA atlas scans with manual segmentations. Each CCTA atlas was registered to the test CCTA using affine and subsequent non-rigid registration with *elastix* [42]. The heart and aorta segmentations were transformed to the test CCTA scan. Segmentations from all atlas scans were combined through majority voting. Subsequently, the LAD, LCX and RCA centerlines in the test CCTA were tracked using an adapted multi-scale vesselness filter [70, 71]. To initialize tracking, the coronary ostia were automatically detected and, in case of failed detection, manually annotated. The heart, ascending aorta and coronary centerline segmentations were transformed from the test CCTA scan to the test CSCT scan in two steps. The test CCTA scan was first registered to the test CSCT scan using affine and subsequent non-rigid registration and the heart segmentation was mapped from the test CCTA onto the test CSCT scan. Then, an additional non-rigid registration was applied, using the transformed heart segmentation in the test CSCT as a mask. Subsequently, the ascending aorta segmentation and coronary artery centerlines were also propagated from the CCTA to the CSCT. A morphological dilation was applied to the coronary centerlines in the test CSCT and the intersection of this dilation and the heart segmentation was considered a coronary artery mask. The aortic segmentation was subtracted from this mask to exclude potential aortic calcifications. CAC was identified using supervised pattern recognition. Only candidate CAC lesions (> 130 HU, < 600 voxels) contained in the coronary artery mask of the CSCT were extracted and described by volume, intensity features, and location in the heart coordinate system. Candidates were classified as CAC or non-CAC by means of a support vector machine trained with the training set provided in the framework. Identified CAC was assigned to an artery using the previously obtained coronary artery labels.

Method C - Siemens Corporate Technology (Princeton, NJ)

CAC was identified by analyzing the test CCTA scan and the test CSCT scan. First, in both the CCTA scan and the CSCT scan the pericardium and aortic root were automatically segmented. In both scans, marginal space learning (MSL) was used to estimate the position, orientation and size of the heart [72]. Mean shapes for the pericardium and aortic root, based on a set of example shapes, were aligned with the estimated pose as initial segmentations. These were then refined using an active shape model. The sternum and ribs were explicitly segmented and subtracted from the obtained pericardium segmentation. LAD, LCX and RCA coronary artery centerlines in the test CCTA scan were automatically extracted by a combination of model-driven and data-driven methods [73]. These centerlines were propagated

from the test CCTA onto the test CSCT, by aligning the coronary root and pericardium contours in both scans using point-wise correspondence. The transformation field within the pericardium was interpolated using a thin-plate-spline model. A coronary artery mask in the CSCT was determined as the intersection of the pericardium segmentation and the coronary artery centerlines, dilated by 1.5 mm. The aortic root was excluded from this mask. Finally, CAC was identified using supervised pattern recognition. Only candidate CAC lesions (> 130 HU, > 3 voxels, 6-connected) that were at least partially in the coronary artery mask of the CSCT were extracted and described with volume, intensity features, distance to the closest centerline and coordinates in a heart coordinate system determined by the pericardium segmentation. Candidates were classified as CAC or non-CAC by means of a random forest classifier trained with the training set provided in the framework. Identified CAC was assigned to an artery using the previously obtained coronary artery labels.

Method D - Konica Minolta Inc. (Osaka, Japan)

CAC was identified by analyzing the test CCTA scan and the test CSCT scan. First, a Gaussian filter was applied to the axial slices in the test CCTA. The lung area in this scan was identified based on HU values and removed. Next, coronary arteries in the test CCTA scan were segmented using region growing. To initialize this process, a point in the ascending aorta was manually identified. A HU threshold was determined based on the intensity of voxels surrounding this point. Subsequent region using this threshold identified connected aortic and coronary voxels. The top-most slice in which the identified component appeared disconnected determined the location of the left coronary trunk. The right coronary trunk was identified using an analogous strategy. Subsequently, a multi-scale vesselness filter was used to enhance vessels in the test CCTA image [70]. The method then extracted the left and right coronary tree separately by region growing based on intensities and the vesselness values. Next, the test CCTA scan was registered to the test CSCT scan using non-rigid registration. The extracted coronary arteries were transformed from the test CCTA onto the test CSCT to create a coronary artery mask in the test CSCT. A CAC candidate mask was generated by thresholding the test CSCT at 130 HU. Finally, voxels in the intersection of the coronary artery mask and the candidate mask were identified as CAC.

Method E - University Medical Center Utrecht (Utrecht, The Netherlands)

CAC was identified by analyzing only the test CSCT scan. First, the position of the LAD, LCX and RCA coronary artery centerlines in the test CSCT scan was estimated. For this, ten CSCT atlas scans with LAD, LCX and RCA centerline annotations were used. The centerlines had previously been manually annotated in corresponding atlas CCTA scans, and transformed to the atlas CSCT scans. Each CSCT atlas scan was registered to the test CSCT scan using affine and subsequent elastic registra-

tion with *elastix* [42]. The obtained transformation was used to map the centerlines from the atlas CSCT onto the test CSCT. Next, for LAD, LCX and RCA separately, a centerline was estimated by iterative fusing of the propagated centerlines. In each iteration, the artery with the largest Fréchet distance to the geometric median was discarded. The final coronary artery estimate was the geometric median of the remaining centerlines after termination. Finally, CAC was identified using supervised pattern recognition. All candidate CAC lesions (> 130 HU, 6-connected, > 1.5 mm^3) in the CSCT were extracted and described by volume, shape features, intensity features and location features based on the estimated coronary artery centerlines. Candidate lesions were classified as CAC in the LAD, LCX or RCA or as non-CAC by a multiclass randomized decision forest trained with 237 previously acquired CSCT scans. This method was developed by the framework organizers, who had access to the reference annotations of all data sets. A detailed description of this method is provided in [15].

3.2.5 Evaluation

CAC identified by the (semi-)automatic methods was compared with the reference standard. In addition, to evaluate the performance of human experts, CAC identified by the two expert observers prior to the consensus reading was compared with the reference standard.

Lesions were defined as 6-connected components of voxels > 130 HU with volume > 1.5 mm^3. The volume of a lesion (in mm^3) was determined as the number of voxels in the lesion multiplied by the voxel volume. The detection of CAC lesions was evaluated using sensitivity, positive predictive value (PPV) and the F_1 score, i.e. the harmonic average of sensitivity and PPV (2·(sensitivity·PPV)/(sensitivity+PPV)), weighted by both the total number and total volume of lesions in the test set. For a statistical pairwise comparison of the performance of methods and observers on the detection of CAC lesions, McNemar's test was performed, with Bonferroni correction for 21 pairwise comparisons so that $\alpha = 0.05/21 = 0.0024$ [74]. The determination of per patient CAC volume was evaluated using the intraclass correlation coefficient (ICC) for absolute agreement with the reference standard.

For each lesion, in addition to CAC volume, the Agatston score was determined as $\sum_{s \in S} a_s \cdot d_s$, where S is the set of image slices containing the lesion, a_s is the total lesion area (in mm^2) in s, and d_s is a density factor determined by the lesion's maximum attenuation in s (130–199 HU: 1, 200–299 HU: 2, 300-399 HU: 3, ≥ 400 HU: 4) [9]. Because the Agatston score is based on images with 3 mm section thickness and increment, a linear correction factor $\frac{2.5}{3.0}$ was applied for GE images with section thickness and increment of 2.5 mm. Per patient Agatston scores were determined by addition of lesion Agatston scores. To compare per patient Agatston scores with the reference standard, Bland-Altman plots were used. In addition, a CVD risk category was computed per patient based on the Agatston scores (I: 0, II: 1-100, III: 101-300, IV: >300) and agreement with the reference standard was computed using linearly

Table 3.3: Sensitivity, positive predictive value (PPV) and F_1 score for the detection of CAC lesions, weighted by number of lesions (Lesions) and by volume in mm^3 (Volume), achieved by each method and the two expert observers. The highest value in each column is shown in boldface, separately for (semi-)automatic methods and observers.

	Sensitivity (%)		PPV (%)		F_1 score (%)	
	Lesions	Volume	Lesions	Volume	Lesions	Volume
Method A[5]	62.3	85.3	87.9	93.7	73.9	89.3
Method B[6]	**94.0**	**98.9**	**95.5**	94.8	**94.7**	**96.8**
Method C[5]	83.9	94.9	95.3	93.8	89.2	94.3
Method D[6]	51.6	57.3	64.7	68.5	57.4	62.4
Method E[5,4]	84.5	93.5	95.0	**95.9**	89.5	94.7
Observer 1	94.3	98.6	**99.3**	**98.6**	96.8	**98.6**
Observer 2	**99.1**	**99.9**	99.1	95.5	**99.1**	97.5

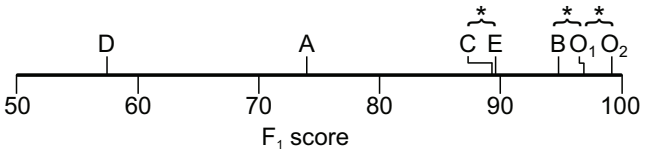


Figure 3.2: F_1 scores for the identification of CAC lesions in the test set, for evaluated methods A, B, C, D and E and the two observers O_1 and O_2. The F_1 score was computed as the harmonic average of sensitivity and PPV ($2 \cdot (\text{sensitivity} \cdot \text{PPV})$), weighted by number of lesions. Pairs indicated by a brace and an asterisk were found not to be statistically different by McNemar's test with Bonferroni correction for 21 comparisons ($\alpha = 0.05/21$). All other pairs performed significantly differently.

weighted Cohen's kappa. Finally, to evaluate per artery CAC volume determination, the ICC for absolute agreement with the reference standard was computed.

The evaluation was implemented in MATLAB (Version R2013a, Mathworks, Natick, MA, USA).

3.3 Results

3.3.1 Detection of CAC lesions

The 40 test exams contained a total number of 316 CAC lesions, corresponding to 9018.5 mm^3. Sensitivity for individual lesion identification ranged from 52% to 94%, corresponding to a sensitivity of 57% to 99% for CAC volume. The PPV ranged from 65% to 96% for lesion identification and from 69% to 96% for CAC volume

[5]Automatic method
[6]Semi-automatic method

Automatic

Reference	FN	LAD	LCX	RCA	Total
FP	-	5	7	15	27
LAD	38	**90**	1	0	129
LCX	27	2	**40**	1	70
RCA	54	0	0	**63**	117
Total	119	97	48	79	

a) Method A

Semi-automatic

Reference	FN	LAD	LCX	RCA	Total
FP	-	3	0	11	14
LAD	7	**119**	3	0	129
LCX	7	1	**57**	5	70
RCA	5	0	0	**112**	117
Total	19	123	60	128	

b) Method B

Automatic

Reference	FN	LAD	LCX	RCA	Total
FP	-	6	0	7	13
LAD	11	**108**	10	0	129
LCX	19	7	**42**	2	70
RCA	21	0	0	**96**	117
Total	51	121	52	105	

c) Method C

Semi-automatic

Reference	FN	LAD	LCX	RCA	Total
FP	-	5	1	83	89
LAD	51	**70**	8	0	129
LCX	32	3	**30**	5	70
RCA	70	0	0	**47**	117
Total	153	78	39	135	

d) Method D

Automatic

Reference	FN	LAD	LCX	RCA	Total
FP	-	6	1	7	14
LAD	10	**109**	10	0	129
LCX	16	2	**51**	1	70
RCA	23	0	0	**94**	117
Total	49	117	62	102	

e) Method E

Manual

Reference	FN	LAD	LCX	RCA	Total
FP	-	1	0	1	2
LAD	5	**122**	2	0	129
LCX	5	1	**64**	0	70
RCA	8	0	7	**102**	117
Total	18	124	71	103	

f) Observer 1

Manual

Reference	FN	LAD	LCX	RCA	Total
FP	-	0	0	3	3
LAD	1	**126**	2	0	129
LCX	1	0	**67**	2	70
RCA	1	0	1	**115**	117
Total	3	126	70	120	

g) Observer 2

Figure 3.3: Confusion matrices for lesion labeling per coronary artery. Each cell shows the number of lesions assigned a coronary artery label in the reference standard (rows) and by the (semi-)automatic methods or observers (columns). The diagonal cells show the number of lesions for which the labeling by the reference standard and (semi-)automatic methods or observers agreed. The off-diagonal cells show the number of lesions where they disagreed. Labels were left anterior descending artery (LAD), left circumflex artery (LCX), right coronary artery (RCA) and negative (Neg). Note that the cells in the first rows and columns show false positive and false negative errors, respectively.

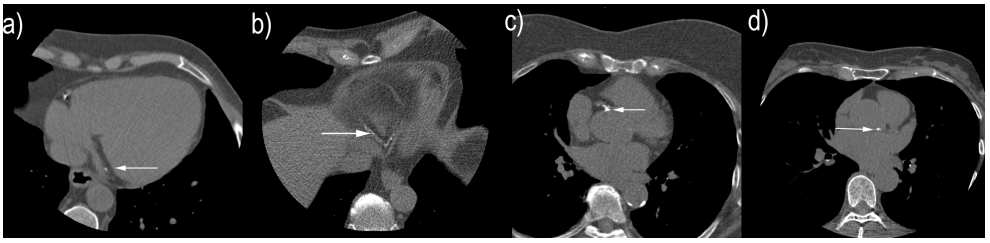


Figure 3.4: Examples of typical false positive and false negative errors made by the evaluated methods. (a) A coronary artery calcification (CAC) lesion in the distal left circumflex artery which was missed by all methods. (b) Two CAC lesions in the distal right coronary artery which were missed by all methods. (c) A large aortic calcification at the right coronary ostium which was incorrectly detected as CAC by three methods. (d) An aortic calcification at the left coronary ostium which was incorrectly detected as CAC by three methods.

Table 3.4: The two-way ICC for absolute agreement with 95% confidence interval between reference and obtained CAC volume per patient (Patient) and per coronary artery (left anterior descending artery [LAD], left circumflex artery [LCX] and right coronary artery [RCA]) shown for each method and for the two expert observers. The highest value in each column is shown in boldface, separately for (semi-)automatic methods and observers.

	Patient	LAD	LCX	RCA
Method A[5]	0.97 (0.94-0.98)	0.98 (0.96-0.99)	0.95 (0.91-0.98)	0.90 (0.81-0.94)
Method B[6]	**0.99 (0.99-1.00)**	**1.00 (1.00-1.00)**	**1.00 (0.99-1.00)**	**0.96 (0.93-0.98)**
Method C[5]	0.98 (0.97-0.99)	0.94 (0.88-0.97)	0.87 (0.77-0.93)	0.96 (0.92-0.98)
Method D[6]	0.60 (0.36-0.76)	0.33 (0.04-0.57)	0.57 (0.32-0.74)	0.51 (0.25-0.71)
Method E[5,4]	0.99 (0.97-0.99)	0.98 (0.97-0.99)	0.98 (0.96-0.99)	**0.96 (0.93-0.98)**
Observer 1	**1.00 (1.00-1.00)**	**1.00 (1.00-1.00)**	0.99 (0.98-0.99)	**0.98 (0.97-0.99)**
Observer 2	0.98 (0.97-1.00)	**1.00 (1.00-1.00)**	**1.00 (1.00-1.00)**	0.90 (0.83-0.95)

identification. The F_1 score ranged from 57% to 95% for lesion identification and from 62% to 97% for CAC volume identification. Detailed results obtained by all methods and by the two expert observers are listed in Table 3.3. For all methods as well as for the experts, the sensitivity was higher for CAC volume than for CAC lesions. This indicates that correctly identified CAC lesions were larger in volume than missed lesions.

Fig. 3.2 shows F_1 scores for lesion identification obtained by the evaluated methods and the two expert observers. All pairwise differences were analyzed using McNemar's test with Bonferroni correction for multiple comparisons. The null hypothesis that a pair of methods or observers performed equally well held for three pairs: methods C and E ($p = 1.0000$), method B and observer 1 ($p = 0.0315$), and the two observers ($p = 0.0094$). All other pairs yielded $p < 0.05/21$; one method or observer in the pair significantly outperformed the other.

Fig. 3.3 shows confusion matrices comparing reference and (semi-)automatically determined labels for all lesions in the test exams. All methods made more false negative (FN) errors than false positive (FP) errors (on average 0.5 – 3.8 FN errors vs. 0.3 – 2.2 FP per patient). However, for all methods FP errors were typically larger than of FN errors (on average 19.2 – 60.8 mm^3 per FP error vs. 5.2 – 10.0 mm^3 per FN error). Fig. 3.4 shows typical FN and FP errors. Eight FN errors were common to all methods. With the exception of one larger lesion in the distal LCX (Fig. 3.4a), these lesions were small in volume, mostly located in the proximal LAD, distal LCX and distal RCA (Fig. 3.4b). No single FP error was common to all methods, but six identical FP errors were made by three methods. Five of these were located at the right coronary ostium (Fig. 3.4c) and one at the left coronary ostium (Fig. 3.4d).

3.3.2 CAC scoring per patient

Median (interquartile range) reference CAC volume per patient was 84.4 (6.6–327.4) mm^3. Table 3.4 lists the two-way ICC for absolute agreement with the reference standard for both the (semi-)automatic methods and the two expert observers. The ICC ranged from 0.60 (0.36–0.76) to 0.99 (0.99–1.00) for the evaluated methods. For the two observers, the ICC values were 1.00 (1.00–1.00) and 0.98 (0.97–1.00), respectively.

Median (interquartile range) reference Agatston score per patient was 72.0 (3.9–361.5). The Bland-Altman plots in Fig. 3.5 show differences between the reference standard and the (semi-)automatic methods, as well as between the reference standard and the two expert observers. Two patients with calcifications close to the coronary ostia and one patient with CAC lesions in the distal parts of the coronary arteries are marked in each plot.

Fig. 3.6 shows confusion matrices for CVD risk based on Agatston scores. Linearly weighted Cohen's kappa ranged from 0.80 to 1.00. The patients assigned to an incorrect risk category were, with one exception, assigned to a neighboring category. Overall, 26/40 (65%) patients were assigned to their correct risk category by all methods.

3.3.3 CAC scoring per artery

The exams in the test set contained 129/316 (41%) LAD, 70/316 (22%) LCX and 117/316 (37%) RCA CAC lesions. The results in Figure 3.3 show that the methods assigned 88%-98% of correctly identified CAC lesions to their reference artery, while this number was 97% and 98% for the observers. Up to 25% of all errors were caused by assigning CAC lesions to the incorrect artery. Fig. 3.7 shows examples of CAC lesions that were assigned to the incorrect coronary artery by multiple methods. CAC lesions in the LAD were incorrectly assigned to the LCX (Fig. 3.7a), while CAC lesions in the LCX were incorrectly assigned to the LAD (Fig. 3.7b). One lesion was labeled as LCX in the reference standard, but identified as RCA by three methods

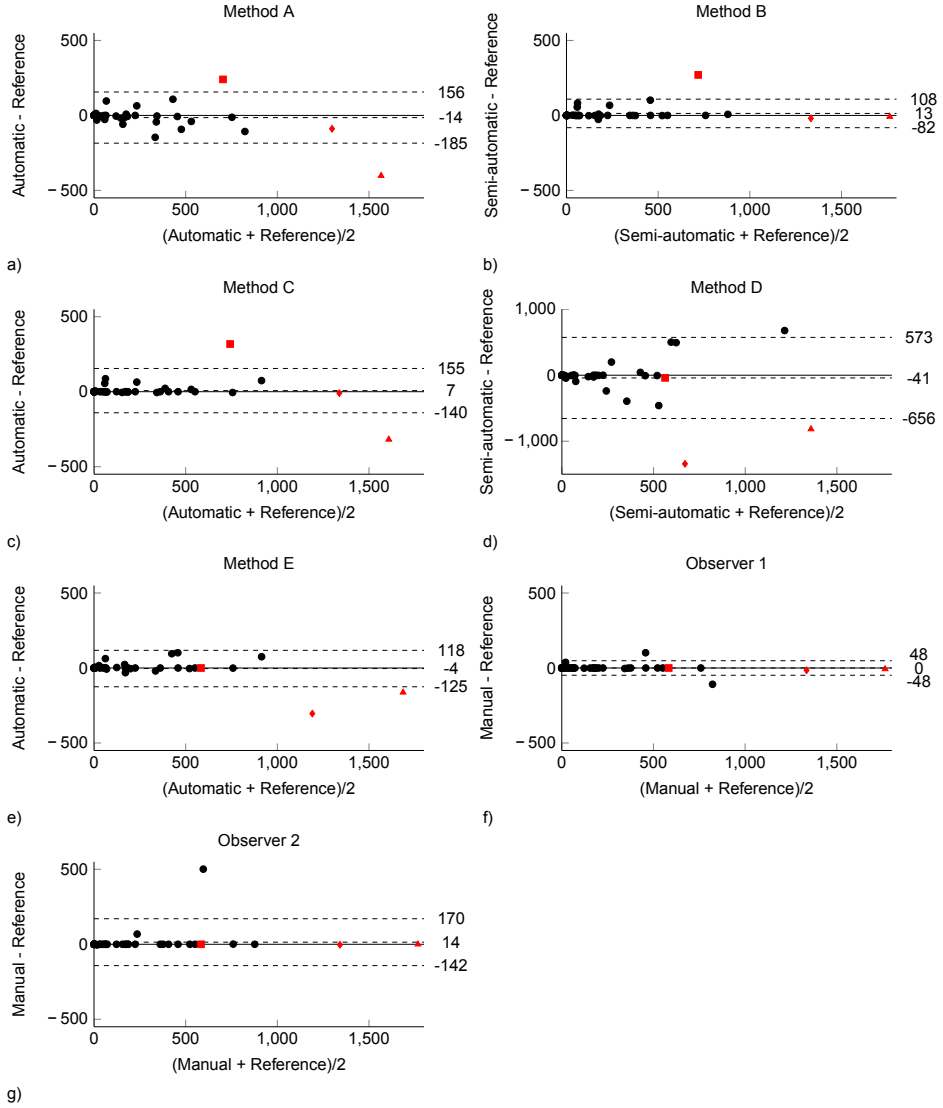


Figure 3.5: Bland-Altman plots comparing per patient reference Agatston scores with (semi-)automatic methods and expert observers. Three patients for whom the reference standard and the evaluated methods showed the largest disagreement are marked in red in all plots: one patient with a large coronary artery calcification (CAC) lesion at the boundary of the ascending aorta and the right coronary artery (RCA, square), one patient with large CAC lesions in the acute marginal branch of the RCA (triangle), and one patient with extensive CAC in the distal left circumflex artery and the RCA (diamond). Dashed lines represent the mean and the limits of agreement of differences (±1.96 SD).

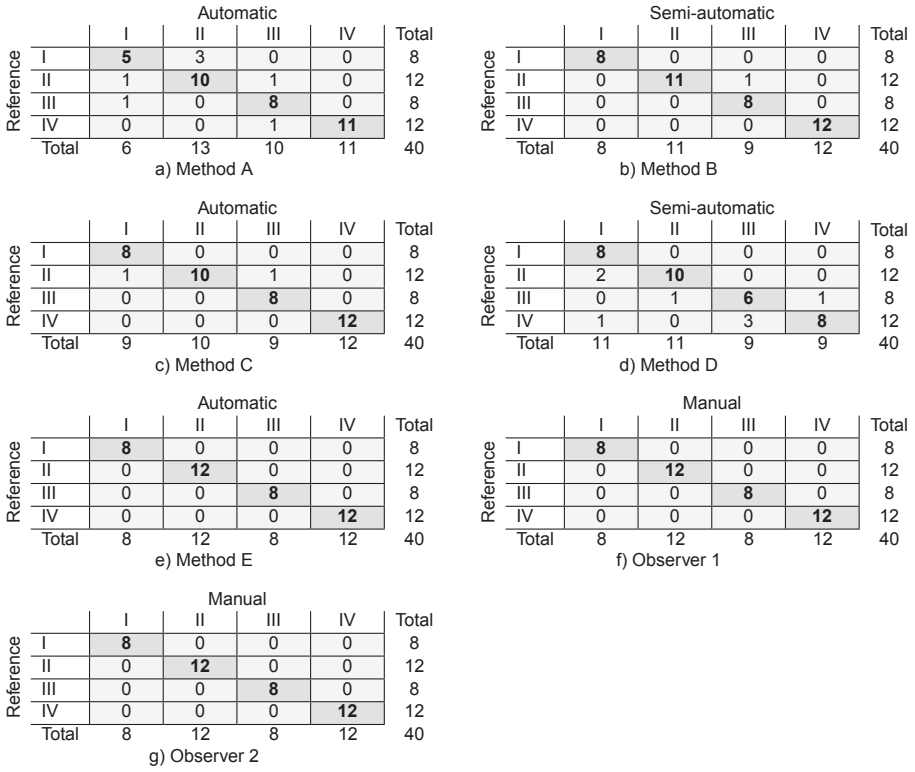


Figure 3.6: Confusion matrices for cardiovascular disease (CVD) risk categorization based on the Agatston score. The diagonal cells show correctly categorized patients. The cells above and below the diagonal show patients whose CVD risk was overestimated or underestimated, respectively. The following linearly weighted Cohen's kappa scores were obtained: (a) 0.88, (b) 0.98, (c) 0.96, (d) 0.80, (e) 1.00, (f) 1.00, and (g) 1.00.

(Fig. 3.7c). Retrospective expert inspection of the reference standard revealed that this lesion was located in the RCA. No labeling errors were made between the LAD and the RCA.

Median (interquartile range) reference CAC volume per artery was 52.0 (7.1–189.2), 1.7 (0.0–105.3) and 6.1 (0.0–69.2) mm^3 for LAD, LCX and RCA, respectively. Table 3.4 lists the two-way ICC for absolute agreement with the reference standard for the (semi-)automatic methods and the two expert observers. Three methods obtained a higher agreement for RCA volume than Observer 2, who incorrectly assigned a large aortic calcification at the right coronary ostium to the RCA.

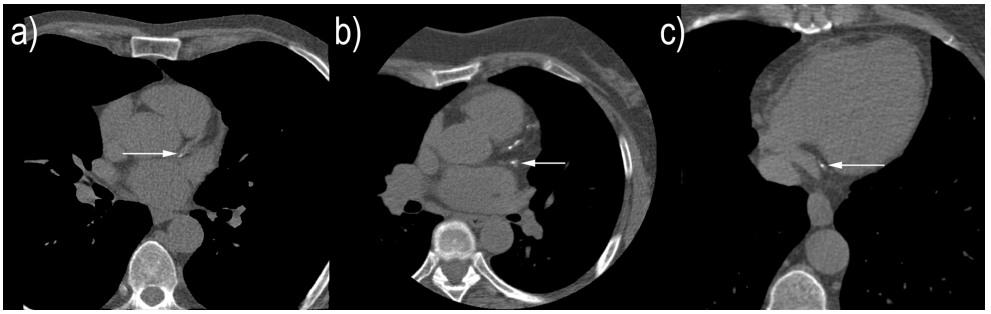


Figure 3.7: Examples of typical labeling errors made by the evaluated methods. (a) A coronary artery calcification (CAC) lesion in the left anterior descending artery (LAD) which was assigned to the left circumflex artery (LCX) by three methods. (b) A CAC lesion in the LCX which was assigned to the LAD by three methods. (c) A CAC lesion which was labeled as LCX in the reference standard, but assigned to the right coronary artery (RCA) by three methods. Retrospective analysis by an expert observer revealed that this lesion was indeed located in the RCA.

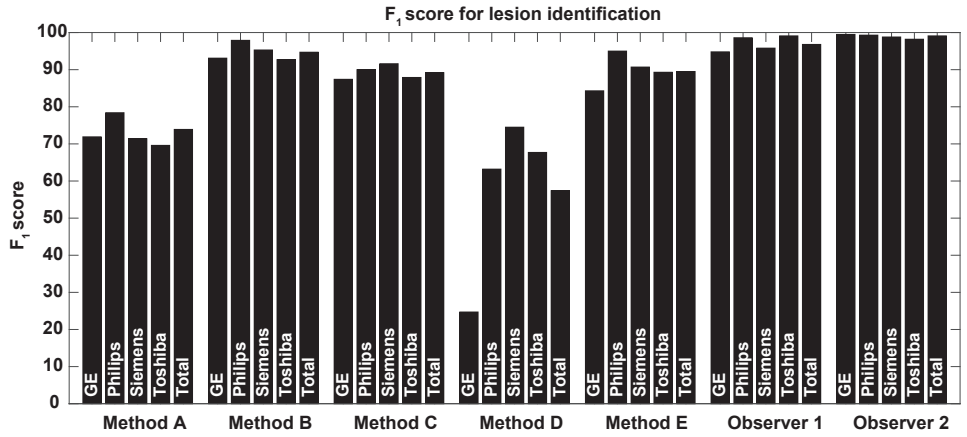


Figure 3.8: F_1 score for lesion identification on scans obtained on four different scanners (GE, Philips, Siemens, Toshiba), and the overall test set (Total), for the evaluated methods A, B, C, D and E, and the two expert observers. The F_1 score was computed as the harmonic average of sensitivity and PPV ($2 \cdot (\text{sensitivity} \cdot \text{PPV})$), weighted by number of lesions.

3.3.4 Performance on scans acquired with different CT scanners

Test exams from the GE, Philips, Siemens and Toshiba scanner contained 102, 72, 86 and 56 CAC lesions with total CAC volumes 2772.8, 3063.0, 1613.2 and 1569.4 mm^3, respectively. Fig. 3.8 shows the F_1 score for lesion identification separately for scans acquired on each scanner, as well as for the total test set, for the evaluated methods A, B, C, D and E and the two expert observers. This provides an indication of the performance of the evaluated methods across scans from different vendors and shows the effect of each scanner on overall performance of the methods. Note that the number of scans included per scanner was small, and it is likely that other factors than only the scanner vendor affected the differences in performance. Hence, this comparison does not provide sufficient statistical power to generalize the results to other data sets or clinical scenarios.

3.4 Discussion

We have presented an independent evaluation framework for (semi-)automatic methods performing coronary calcium scoring in cardiac CT. The framework provides clinical exams consisting of a CSCT scan with a corresponding CCTA scan and standardized criteria for performance evaluation. Five methods were evaluated within this framework. All methods achieved linearly weighted Cohen's kappa ≥ 0.80 with the reference standard for patient CVD risk categorization. On a lesion level, some methods can identify CAC in CSCT with a sensitivity and positive predictive value close to that of expert observers. Nevertheless, all evaluated methods made common errors, typically at the coronary ostia and in distal segments of the coronary arteries.

Three of the evaluated methods were fully automatic, requiring no user interaction. The other two methods were semi-automatic, requiring limited user interaction to manually initialize or correct automatic coronary artery extraction. Although this level of expert interaction is substantially lower than in manual CAC scoring as performed in clinical settings, it might reduce the potential of these methods for large studies or screening settings [75]. The focus of the evaluation criteria was on the accuracy of CAC identification and not on time required for the analysis; hence methods were generally not optimized for speed. Consequently, the automatic methods took up to 80 minutes, which exceeds the time manual scoring would take. However, these methods required no interaction and CAC scores may thus be obtained 'off-line' without additional workload for the clinician.

Because CSCT is the standard for CAC scoring in the clinic, methods were evaluated on their ability to identify CAC in CSCT. However, for each CSCT, a corresponding CCTA scan was provided. Four methods (A, B, C, D) used this CCTA scan in addition to the CSCT. In clinical practice, determining a CAC score based on only CSCT could be advantageous, as CCTA is not always acquired in conjunction with CSCT [76]. However, information in the CCTA could improve the identification of CAC, due to increased visibility of the coronary arteries in CCTA compared

to CSCT. Hence, ideally, a method would be able to process stand-alone CSCTs, as well as pairs of CSCT and CCTA scans when both are available. Method A used the test CCTA as an intermediate for registration of the test CSCT and their atlas CCTA and could be adapted to use only test CSCT, as demonstrated in previous work [13]. Conversely, method E, which now used only CSCT, could be extended to also analyze CCTA. Methods B, C, and D relied on coronary artery segmentation in the patient CCTA, which might limit their performance when no CCTA is available.

Clinical CAC scoring commonly provides a single measure for patient CAC burden. However, it has been shown that the distribution of CAC over arteries is associated with total coronary plaque burden [77] and is highly predictive of CVD events [78] or future coronary revascularization [79]. Hence, each method's ability to determine this distribution was evaluated. The methods assigned the vast majority (88%-98%) of correctly identified CAC lesions to the correct artery. The evaluated methods identified CAC lesions in the LAD most accurately, followed by those in the LCX and RCA. Reduced performance in the RCA might be explained by the presence of CAC in distal segments, which is more common in the RCA than in the LAD and LCX [78]. The identification of distal CAC in the RCA might be challenging for artery-tracking methods such as B, C and D, which commonly extract proximal parts of the coronary artery better than distal parts [54], as well as for multi-atlas based methods such as A and E, which are sensitive to anatomical variations in the distal coronary arteries. In addition, differences between identification of CAC in the LAD and in the LCX or RCA might be explained by the final supervised classification step common to methods A, B, C and E, which requires a sufficient number of training samples. The 32 provided training CTs contained more LAD than LCX or RCA CAC lesions, which potentially limits each method's ability to distinguish CAC lesions in the LCX and RCA.

The provided CTs originated from four different CT scanners. Even though scanning protocols for CSCT are standardized [11], CTs acquired with different scanners have been shown to yield significantly different CAC scores [56]. The results showed that methods A, B, C and E obtained similar F_1 scores for CTs originating from different scanners. Methods A, B and E were developed with CTs that were not provided in the framework, acquired with Philips (method E) or Siemens (methods A and B) CT scanners. This could potentially lead to a positive bias towards CTs from the same scanner vendor. However, the results showed only a slightly positive bias toward Philips for method E and even a slightly negative bias toward Siemens for methods A and B. Hence, the results did not indicate a bias towards the scanners used in method development. Method D failed to identify CAC in two scans acquired with the GE scanner, and made several large FP errors on another scan acquired with the GE scanner. Hence, its performance on scans obtained with the GE scanner was substantially lower than on scans obtained with the other scanners. This occurred due to failure to detect the left and right coronary ostia, and hence failure to extract the coronary artery tree. The extent of differences between scanners, the

limited number of test exams and other factors such as patient characteristics and patient set-up during scanning prohibit us to make conclusions. In future work, the provided CTs could be supplemented with a large number of scans originating from a range of different scanners and scanner vendors allowing an in-depth analysis of the performance of evaluated methods on diverse data.

Large-scale clinical trials have shown that asymptomatic patients with zero CAC have an excellent CVD prognosis, and that the identification of these patients might prevent unnecessary healthcare costs [80]. The evaluated methods were able to differentiate between zero CAC and non-zero CAC patients. Four out of five methods correctly identified all eight patients without CAC, while the fifth method overestimated CAC in three of these patients owing to small FP errors. Conversely, two methods correctly identified all 32 patients with positive CAC, while the remaining three methods incorrectly assigned a zero CAC score to at most three patients. These results indicate that (semi-)automatic CAC scoring might be used to identify zero CAC patients.

Although CAC is typically identified and quantified in dedicated cardiac CT scans, it may also be scored in other CT scans visualizing the heart. This is especially interesting in non-ECG-triggered low-dose chest CT acquired in lung cancer screening. However, automatic CAC scoring in chest CT images poses different challenges, because these scans may contain severe motion artifacts and high levels of noise. Hence, dedicated methods have been developed for this purpose [14, 37] and it is not likely that the methods evaluated in this study would be directly applicable to chest CT. The methods would need to be adjusted to low-dose chest CT, e.g. the requirement of additional CCTA scans for methods A, B, C and D should be removed.

Scans included in this study provide a set of representative cardiac CT scans. Patients were balanced over gender, CVD risk and scanning site. Scans with anatomical abnormalities, intracoronary stents, metal implants severe motion artifacts or extremely high levels of noise determined by visual inspection were excluded. Even though calcium scores determined in such images may not be reliable or scoring might not be possible, in future work the provided CTs could be supplemented with exams containing such abnormalities, to evaluate whether methods would be able to recognize such scans.

3.5 Conclusion

A publicly available standardized framework for the evaluation of (semi)automatic methods for CAC identification in cardiac CT is described. An evaluation of five (semi-)automatic methods within this framework shows that automatic per patient CVD risk categorization is feasible. CAC lesions at ambiguous locations such as the coronary ostia remain challenging, but their detection had limited impact on CVD risk determination.

Chapter 4

2D image classification for 3D anatomy localization: employing deep convolutional neural networks

Based on:
B.D. de Vos, J.M. Wolterink, P.A. de Jong, M.A. Viergever, I. Išgum. "2D image classification for 3D anatomy localization: employing deep convolutional neural networks," *SPIE Medical Imaging*, 2016, 97841Y

Abstract

Localization of anatomical regions of interest (ROIs) is a preprocessing step in many medical image analysis tasks. While trivial for humans, it is complex for automatic methods. Classic machine learning approaches require the challenge of handcrafting features to describe differences between ROIs and background. Deep convolutional neural networks (CNNs) alleviate this by automatically finding hierarchical feature representations from raw images. We employ this trait to detect anatomical ROIs in 2D image slices in order to localize them in 3D.

In 100 low-dose non-contrast enhanced non-ECG synchronized screening chest CT scans, a reference standard was defined by manually delineating rectangular bounding boxes around three anatomical ROIs – heart, aortic arch, and descending aorta. Every anatomical ROI was automatically identified using a combination of three CNNs, each analyzing one orthogonal image plane. While single CNNs predicted presence or absence of a specific ROI in the given plane, the combination of their results provided a 3D bounding box around it.

Classification performance of each CNN, expressed in area under the receiver operating characteristic curve, was $\geq$0.988. Additionally, the performance of ROI localization was evaluated. Median Dice scores for automatically determined bounding boxes around the heart, aortic arch, and descending aorta were 0.89, 0.70, and 0.85 respectively.

The results demonstrate that accurate automatic 3D localization of anatomical structures by CNN-based 2D image classification is feasible.

4.1 Introduction

Localization of specific organs or anatomical regions of interest (ROIs) in medical images is a prerequisite for many tasks in medical image analysis. In segmentation tasks localization is often the first step in reducing the search space in an image. As it is often used as a preprocessing step [81, 82, 83, 84, 85, 86], localization of an anatomy of interest needs to be robust and fast. Localization is considered a trivial task for humans, but it can be a challenge for automatic methods. Standard pattern recognition approaches were successfully employed for organ detection and localization in medical images [82, 87, 88, 89]. Nevertheless, these approaches require profound knowledge of the task at hand. Characteristics differentiating target ROIs from the background have to be translated into numeric representations (i.e. features) that provide the basis for class separation in feature space. Finding the appropriate description of e.g. textures or shapes can be challenging, especially in the presence of pathology.

In recent years, deep convolutional neural networks (CNNs) have gained popularity in image analysis and object detection. In contrast to most machine learning techniques, CNNs take raw images as input and extract hierarchical feature representations themselves, thus eliminating the feature design step. This ability enables CNNs to discriminate a wide variety of different images into correct classes[90].

CNNs were previously used for object localization in 2D natural images by classification of its subimages. For example, Sermanet et al. [91] used a sliding window detector over multiple locations and at multiple scales, essentially performing an exhaustive search. Girshick et al. [92] narrowed the search space by segmentation of the possible target objects using selective search [93].

The aim of this work is to employ the traits of CNNs to localize anatomical ROIs in 3D medical images. The novelty in this approach is that we pose 3D organ detection as a 2D problem. Instead of extracting 2D or 3D subimages, we analyze full 2D image slices and detect presence of a ROI. A combination of CNNs, each classifying one of the three orthogonal image planes, can be used to detect presence of a ROI in extracted 2D image slices. In an unseen image, a ROI can then be localized in 3D by combining the probabilities obtained by the separate CNNs.

4.2 Data

This study includes 100 consecutively selected low-dose non-contrast enhanced non-ECG synchronized chest CT scans, acquired at the University Medical Center Utrecht (Utrecht, The Netherlands) using two CT scanners: a Mx8000 IDT or a Brilliance 16P (Philips Healthcare, Best, The Netherlands). Axial images were acquired at inspiration with 1 mm slice thickness and 0.7 mm increment. The field of view included the outer rib margin at the widest dimension of the thorax. In-plane resolution ranged from 0.55 to 0.87 mm. Scans were reconstructed with moderately smooth kernels (Philips B).

Rectangular bounding boxes were manually drawn around the heart, the aortic arch, and the descending aorta defining the reference volumes of interest in each CT scan. The bounding box around the heart was annotated in the axial plane from the bifurcation of the pulmonary artery to the heart's apex; in the sagittal and coronal planes from its base to its apex, which were clearly visible. The bounding box around the aortic arch was defined in the sagittal and axial view as a section of the aorta connecting the ascending and the descending aorta. The coronal view was used to ensure that the complete arch was included in the boundaries. The bounding box describing the descending aorta was defined using the axial, coronal, and sagittal views to include the part starting from the aortic arch to the level of the heart's apex.

4.3 Method

For each anatomical ROI, three independent CNNs were trained. The CNNs act as ROI detectors, each classifying 2D image slices from one of three orthogonal image planes (axial, sagittal, or coronal). By combining the results in all three image planes, a rectangular 3D bounding box defining the location of the anatomical ROI can be determined.

The dataset was randomly divided into a training and a test set, with equal numbers of subjects. Each of the three CNNs was trained separately using axial, coronal or sagittal image slices. Training sets were randomly divided into two sets: training sets used to train the CNNs, and validation sets used to determine the optimal network and parameter settings. The validation sets contained 25% of the total number of training slices, while preserving prior class label probabilities. Slices labeled as positive were those that contained the anatomical ROI, i.e. those that were inside the reference bounding box. The remaining slices were labeled as negative.

During testing, image slices were extracted from an image in sequential order and presented to the CNN analyzing the corresponding image planes, i.e. determining presence of a ROI. To obtain 3D bounding boxes from the posterior probabilities, a threshold of 0.5 was applied. Thereafter, the largest connected 1D component was retained in each image plane. Finally, a 3D bounding box was defined by broadcasting the binary results to a 3D space and performing voxel-wise multiplication.

Classification was performed by AlexNet [90] CNNs. AlexNet is a deep learning network that has 60 million parameters with 650,000 nodes. It consists of five convolutional layers and three fully connected layers, and produces a softmax output. Each CNN was trained with the NVIDIA Deep Learning GPU Training System (DIGITS) and Caffe [94] in 30 epochs using stochastic gradient descent with a momentum of 0.9, a base learning rate of 0.01, reducing the learning rate by 0.1 every 10 epochs.

The software required 8-bit input images (i.e. 256 different values). Because a CT-scan contains CT numbers ranging from -1000 up to 3095 Hounsfield units (HU)

Table 4.1: Area's under the ROC curves achieved by each CNN.

	Axial	Coronal	Sagittal
Heart	0.988	0.994	0.998
Aortic arch	0.988	0.997	0.992
Descending aorta	0.996	0.996	0.997

and the anatomical regions of interest are mostly soft tissue, prior to training and classification CT numbers in the image slices were set to a window width of 750 and a level of 90 Hounsfield units (HU), and subsequently scaled to an 8-bit grayscale representation. Given that the CNN requires images of fixed size, slices were resized to an isotropic resolution of 1.5 mm per pixel using nearest neighbour interpolation, and were thereafter 0-padded and/or cropped along the outer image borders.

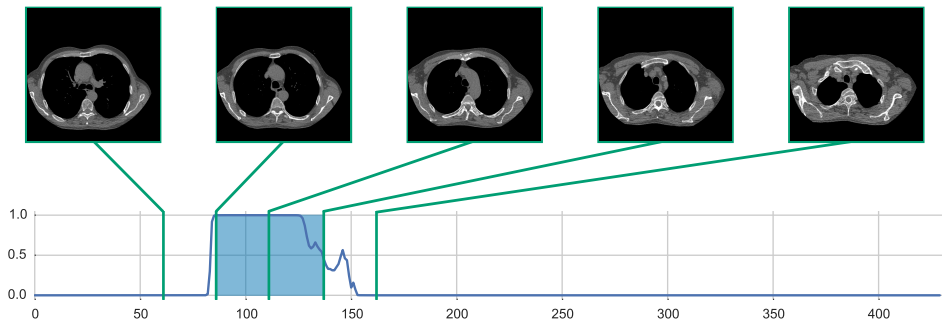
4.4 Results and discussion

An example result is shown in Fig. 4.1. Each CNN determines a probability that an image slice contains a ROI. Therefore, performance of each CNN can be evaluated by a receiver operating characteristics (ROC) curve. Obtained areas under the ROC curves (AUCs) are listed in Table 4.1. The AUCs show that the CNNs can detect presence of anatomical ROIs with high accuracy.

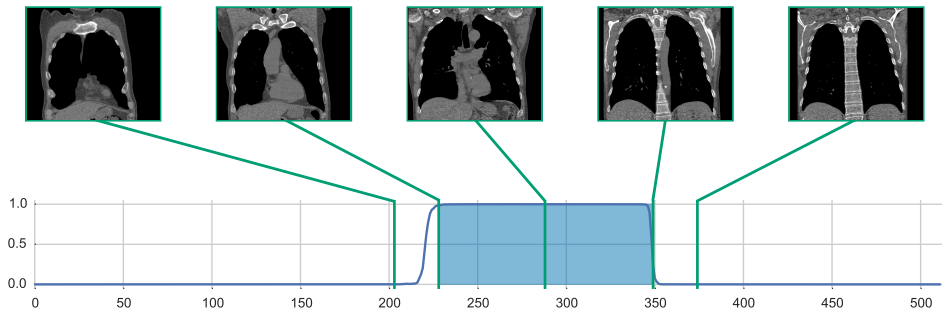
The ability of the method to create a bounding box around anatomical ROIs was evaluated for each image plane and in 3D. For this purpose Dice coefficients were calculated between reference and automatically determined bounding boxes. Moreover, distances between reference and automatic bounding boxes' centroids, and mean distances between the bounding boxes' walls were computed. The obtained results are shown in Fig. 4.2, Fig. 4.3, and Fig. 4.4, respectively.

The highest Dice coefficients (Figure 4.2) were achieved for the heart in the sagittal slices (median 0.98), while the lowest scores were obtained in the axial slices of the aortic arch (median 0.86). Distance in estimated centroids and average wall distances are similar across all detected structures. Smallest distances are seen in sagittal planes, except in sagittal localization of the aortic arch. This is likely caused by arbitrariness of the aortic arch definition and anatomical variation of the arch in that plane. Overall performance for localization of all anatomical ROIs in the axial plane is somewhat lower than in the other planes. This might have been caused by larger variation of anatomy among subjects in this plane compared to the other planes.

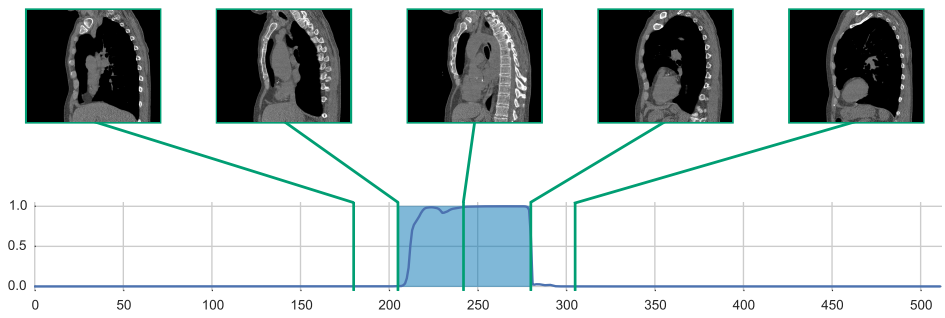
One scan was excluded from further analysis of the aortic arch because the posterior probability did not exceed the detection threshold in any of the axial slices. Therefore, a 3D bounding box could not be defined. Visual inspection showed that the diameter of the aortic arch in this scan was unusually large and the arch was



(a) Axial

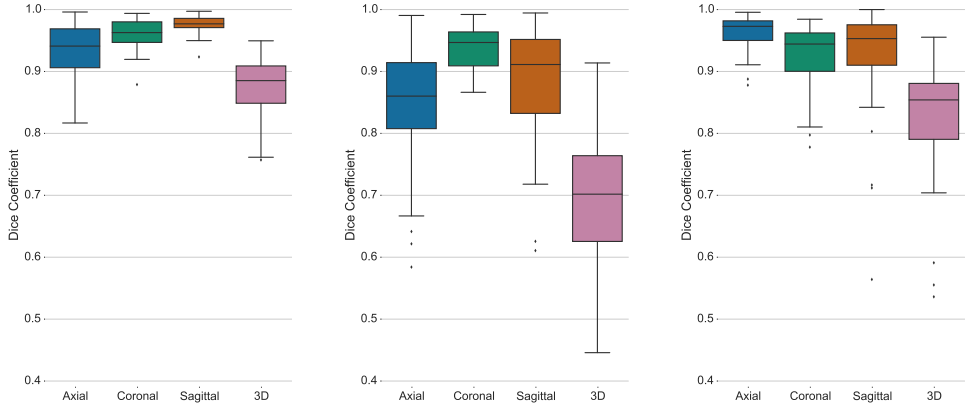


(b) Coronal



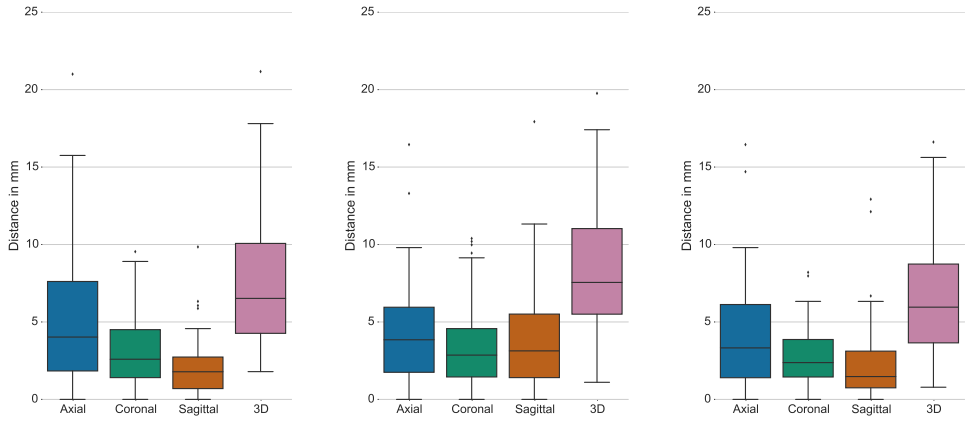
(c) Sagittal

Figure 4.1: Example of aortic arch detection in a chest CT-scan. Results of CNN classification of the (a) axial, (b) coronal, and (c) sagittal image slices. Each subfigure shows five example slices with slices spatially positioned in the scan as indicated by the green lines in the graphs at the bottom of each subfigure. Posterior probabilities for presence of the anatomical ROI are obtained for each slice. In the graphs the posterior probabilities (y-axis) are plotted against slice-numbers (x-axis); the blue opaque areas indicate the reference ROI.



(a) Heart **(b)** Aortic arch **(c)** Descending aorta

Figure 4.2: Dice scores of bounding boxes around the heart, aortic arch, and descending aorta.

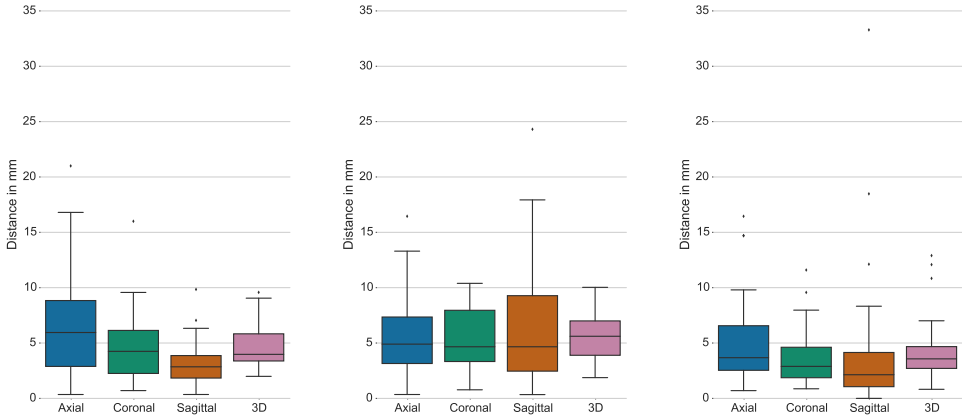


(a) Heart **(b)** Aortic Arch **(c)** Descending Aorta

Figure 4.3: Distance between the automatic and reference bounding box centroids.

situated more superior than in other subjects. Hence, failure of automatic detection might have been caused by the limited number of (representative) training samples.

The here used CNNs were designed to provide a probability per image slice, and are not designed to directly predict a bounding box in 3D image space. Furthermore, per-plane Dice and centroid distance errors accumulate when constructing a 3D bounding box from the per-plane results. Hence, the constructed 3D bounding boxes show increased errors in comparison with performance in separate image planes. Even though a CNN dedicated to 3D organ detection might improve perfor-



(a) Heart **(b)** Aortic Arch **(c)** Descending Aorta

Figure 4.4: Average distance between the automatic and reference bounding box walls.

mance, the obtained results provide a sufficiently accurate localization of the ROIs in 3D by utilizing a readily available CNN analyzing 2D slices.

The method was trained with a limited number of training samples (50) and obtains a bounding box in all scans within 10 seconds. Average distance of all bounding box faces to the reference bounding box faces is 4.8 mm with a standard deviation of 5.2 mm. Criminisi et al. [88] described localization of anatomical structures in CT using a regression forest. That system achieved an average (standard deviation) distance of all reference bounding box faces to the automatic ones of 13.5 mm (13.0 mm). Given that that work analyzed different data and performed localization of different anatomical structures, the results can not be directly compared, but can only be used as an indication.

Unlike specific segmentation tasks where the exact boundaries of the segmentation target are clearly defined, precise definition of a rectangular bounding box around an anatomical structure is challenging. Even though the protocol for setting the reference standard was specified, determining these exact demarcations of the bounding boxes (e.g. the slice in which the cardiac apex starts or the exact slice where the aortic arch stops) is not trivial, especially in non-contrast enhanced low-dose chest CT scans. Visual inspection of the results suggests that most errors obtained by the automatic method might be similar to intra- and inter-observer variations. In our future work, this will be evaluated and compared to automatic performance.

4.5 Conclusions

An automatic method for the localization of anatomical ROIs in low-dose non-contrast enhanced non-ECG synchronized chest CT scans was presented. Localization of anatomical ROIs was performed by combining the output of three CNNs, each analyzing one image plane. The method showed good performance for localization of the heart, the descending aorta, and the aortic arch. The results demonstrate that 3D localization of anatomical ROIs in medical images is feasible by CNNs classifying 2D image slices. The localization method can provide a starting point for complex image analysis tasks.

Chapter 5

Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks

Based on:
J.M. Wolterink, T. Leiner, B.D. de Vos, R.W. van Hamersvelt, M.A. Viergever, I. Išgum. "Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks," *Medical Image Analysis*, 2016, vol. 34, pp. 123–136

Abstract

The amount of coronary artery calcification (CAC) is a strong and independent predictor of cardiovascular events. CAC is clinically quantified in cardiac calcium scoring CT (CSCT), but it has been shown that cardiac CT angiography (CCTA) may also be used for this purpose. We present a method for automatic CAC quantification in CCTA. This method uses supervised learning to directly identify and quantify CAC without a need for coronary artery extraction commonly used in existing methods.

The study included cardiac CT exams of 250 patients for whom both a CCTA and a CSCT scan were available. The volume-of-interest is restricted to an automatically determined bounding box around the heart. The bounding box detection algorithm uses a combination of three convolutional neural networks (CNNs) that each detect the heart in a different orthogonal plane (axial, sagittal, coronal). These CNNs were trained using 50 cardiac CT exams. In the remaining 200 exams, a reference standard for CAC was defined in CSCT and CCTA. Out of these, 100 CCTA scans were used for training, and the remaining 100 for evaluation of a voxel classification method. The method uses ConvPairs, pairs of CNNs. The first CNN in a pair identifies voxels likely to be CAC, and discards the majority of non-CAC-like voxels such as lung and fatty tissue. The second CNN in the pair classifies the identified CAC-like voxels to distinguish between CAC and CAC-like negatives. Given the different task of each CNN, they share their architecture, but not their weights. Input patches are 2.5D or 3D and the CNNs are purely convolutional, i.e. no pooling layers are present and all layers are implemented as convolutions, thereby allowing efficient voxel classification.

The performance of individual 2.5D and 3D ConvPairs with input sizes of 15 and 25 voxels, as well as the performance of ensembles of these ConvPairs, were evaluated by a comparison with reference annotations in CCTA and CSCT. In all cases, ensembles of ConvPairs outperformed their individual members. The best performing individual ConvPair detected 72% of lesions in the test set, with on average 0.85 false positive (FP) errors per scan. The best performing ensemble combined all ConvPairs for a sensitivity of 71% at 0.48 false positive (FP) error per scan. Agreement of this ensemble with the reference mass score in CSCT was excellent (ICC 0.944 [0.918–0.962]). Aditionally, based on the Agatston score in CCTA, this ensemble assigned 84% of patients to the same cardiovascular risk category as reference CSCT.

In conclusion, CAC can be accurately automatically identified and quantified in CCTA using the proposed method. This might obviate the need for a dedicated CSCT scan for CAC scoring, which is regularly acquired prior to a CCTA, and thus reduce CT radiation dose received by patients.

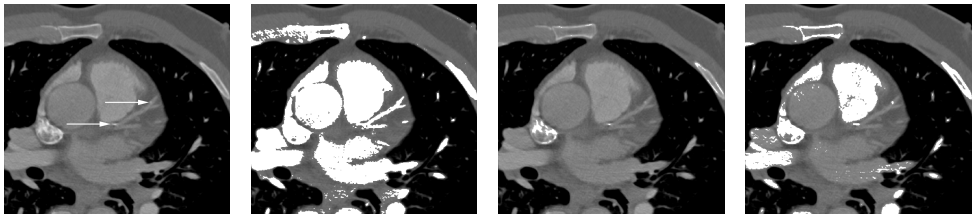
5.1 Introduction

Cardiovascular disease (CVD) is the global leading cause of death. The amount of coronary artery calcification (CAC) as quantified in cardiac CT – the calcium score – is a strong and independent predictor of CVD events [8].

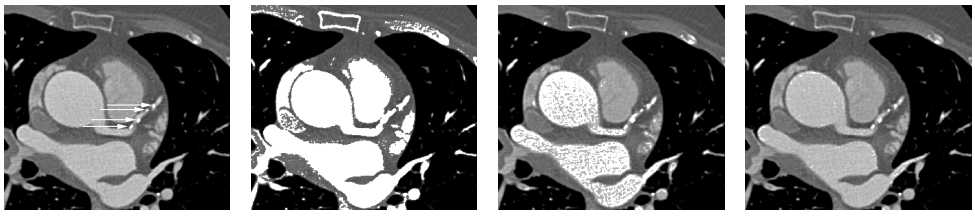
In a clinical cardiac CT exam, a calcium scoring CT (CSCT) scan and a coronary CT angiography (CCTA) scan are typically both acquired. The CCTA scan is used for stenosis detection or identification of non-calcified plaque, and the CSCT scan is used to determine the calcium score [95]. However, it has been shown that CAC may also be quantified in CCTA. In a study by Pavit et al. [19], 85% of patients with a high calcium score in CSCT also had a high calcium score in CCTA (specificity 99%). Moreover, Mylonas et al. [20] showed excellent agreement between CVD risk categories based on calcium scoring in CCTA and categories based on calcium scoring in CSCT (Cohen's linearly weighted $\kappa = 0.93$). A recent survey reported typical radiation doses of 1 mSv for CAC scoring in CSCT [96], while modern techniques allow CCTA acquisitions with 1.5 mSv radiation dose [97]. Hence, performing calcium scoring in CCTA and omitting acquisition of the CSCT scan could reduce the radiation dose of a cardiac CT examination by 40-50% [98].

In clinical practice, CAC is standardly quantified in CSCT by manual identification of groups of connected voxels in the coronary artery that are above a 130 HU threshold and subsequent automatic 3D region growing [9]. This procedure is not applicable to CCTA, due to intravascular contrast material that typically enhances the arterial lumen well beyond 130 HU (Figs. 5.1b, 5.1f). Hence, higher global detection thresholds, ranging from 320 HU [99] to 600 HU [100] have been proposed to emulate CAC scoring in CCTA. However, these fixed thresholds do not consider variations in lumen attenuation in CCTA, which might occur depending on protocols, scanners or contrast agents (Figs. 5.1c, 5.1g). This variation can be taken into account by using patient-specific or scan-specific attenuation thresholds, based on HU values taken from a ROI in the ascending aorta [20] or the proximal coronary arteries [19] (Figs. 5.1d, 5.1h).

Manual identification of CAC in cardiac CT requires substantial expert interaction, which makes it time-consuming and infeasible for large-scale or epidemiological studies. To overcome these limitations, (semi)-automatic calcium scoring methods have been proposed for CSCT (see e.g. [33, 34, 13, 15] and [60]). Wolterink et al. [101] provide a comparison of (semi-)automatic methods for calcium scoring in cardiac CT exams. Similarly, methods have been developed for automatic calcium scoring in CCTA. These methods typically require a (semi)-automatically extracted segmentation of the coronary arteries. Based on this segmentation, CAC has been identified as deviation from a trend line through the lumen intensity [67, 63], as voxels in the extracted arteries with intensities above a patient-specific HU threshold [102], or as deviations from a model of non-calcified artery segments [64]. Mittal et al. [65] did not use a model or threshold to identify CAC, but trained classifiers to identify CAC lesions along an extracted coronary artery centerline. Coronary artery



(a) CCTA scan **(b)** θ_{man}=130 HU **(c)** θ_{man}=600 HU **(d)** θ_{man}=284 HU



(e) CCTA scan **(f)** θ_{man}=130 HU **(g)** θ_{man}=600 HU **(h)** θ_{man}=767 HU

Figure 5.1: Manual CAC identification in CCTA using diverse thresholds (θ_{man}). (a), (e) Two example CCTA images with CAC indicated by white arrows. Voxels with attenuation $> \theta_{man}$ shown in white. (b), (f) θ_{man} = 130 HU [9] oversegments CAC in both images. (c),(g) θ_{man} = 600 HU [100] misses one CAC lesion in (c) and oversegments CAC in (g). (d), (h) a patient-specific threshold based on attenuation in the ascending aorta [20] identifies the individual CAC lesions. Window and level are the same for all images.

tree extraction methods generally show good performance, but they have been reported to fail in patients with complex anatomy, in the distal segments of the coronary arteries, in scans with motion or noise artifacts and in scans with occlusions in the coronary arteries. In addition, severe CAC deposits affect the performance of artery extraction algorithms, restricting their applicability in CAC identification [54]. Manual correction of incorrectly segmented coronary arteries is often time-consuming and tedious.

We propose identification of CAC without initial coronary artery tree extraction. In contrast to previously proposed methods, our algorithm uses supervised learning to directly identify CAC in CCTA. Supervised learning using nearest-neighbor, SVM and randomized decision tree classifiers has been previously applied to CAC identification in CSCT (e.g. [14, 62, 13]). However, these methods cannot be applied in CCTA, as they classify potential CAC lesions, extracted using a clinical 130 HU threshold. In CCTA, it is non-trivial to distinguish between CAC and attenuated lumen, and the application of a predefined single detection threshold to extract potential CAC lesions is not feasible. Instead, the proposed method identifies CAC voxels to segment lesions.

CAC voxel identification in CCTA is a challenging and extremely unbalanced classification problem. The proposed algorithm therefore first limits the volume-of-

interest (VOI) to a bounding box around the heart, extracted using our previously proposed algorithm [103]. Thereafter, voxels in this VOI are classified using convolutional neural networks (CNNs). Recently, CNNs have been successfully used in natural image classification, image segmentation and object detection. In addition, they have been used in several medical image analysis tasks, e.g. knee cartilage segmentation [104] lymph node detection [105], brain tissue segmentation [106], and pulmonary nodule classification [30]. In the proposed algorithm, CNNs automatically extract texture features from triplanar 2.5D or volumetric 3D input samples, which are combined with spatial features derived from a normalized coordinate system defined in the VOI. To classify voxels as CAC or non-CAC, a pair of CNNs is used. These CNNs are linked by training and together are called a ConvPair. The first CNN identifies voxels likely to be CAC. Such voxels are further classified by the second CNN, which distinguishes between CAC and CAC-like negatives. We propose a purely convolutional CNN architecture, which allows for fast evaluation times and can be directly applied to arbitrarily sized CCTA images. In addition, we present experiments showing that combinations of different architectures can achieve higher CAC identification performance than individual architectures.

We have previously proposed a method for CAC scoring in CCTA using a combination of a CNN and a Random Forest classifier [62]. This work extends our previous work in several ways. First, the classification procedure has been modified. Our previously proposed method used a CNN for voxel classification and a Random Forest classifier for lesion classification. The current method uses two sequential CNNs for voxel classification. Second, in our previous work, candidate voxels for classification were selected based on the image intensity histogram. In the current work, we classify all voxels within the VOI, regardless of intensity, hence no assumptions are made about CAC HU values. Third, location features were previously extracted using a time-consuming elastic registration preprocessing step. In the current method, this registration step is omitted in favor of our very fast CNN-based bounding box detection technique [103]. Fourth, in our previous work we only evaluated triplanar 2.5D input with one input size. In the current work, we provide a comparison between 2.5D and volumetric 3D input, between input with different sizes, as well as experiments with ensembles combining these input representations. Fifth, the CNN architecture in our previous work required a time-consuming scan algorithm with many redundant operations for neighboring candidates. Here, we use a purely convolutional network for efficient voxel classification. Finally, in this work an evaluation on a substantially larger set of scans has been performed, and a thorough comparison with clinically used CSCT CAC scores, as well as interobserver variability, are provided.

5.2 Data

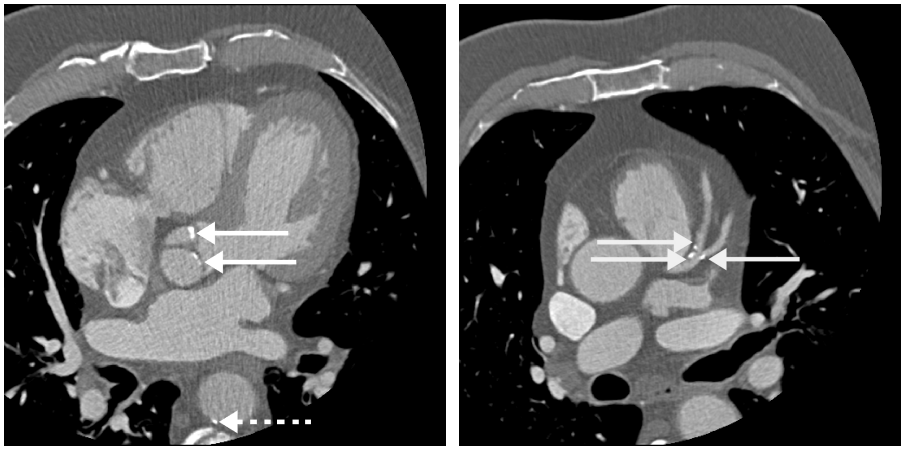
In this study, clinically obtained cardiac CT exams were included of 250 consecutively scanned patients. Each exam consists of a CSCT and a CCTA scan, made on a 256-detector row scanner (Philips Brilliance iCT, Philips Medical, Best, The Netherlands). The CSCT scans were acquired using a standard calcium scoring protocol with 120 kVp tube voltage and 55 mAs tube current, with ECG-triggering and without contrast enhancement. Reconstructed sections had 3.0 mm spacing and thickness. The CCTA scans were acquired with 120 kVp tube voltage and 210-300 mAs tube current, with ECG-triggering and contrast enhancement. Reconstructed sections had 0.45 mm spacing and 0.90 mm thickness. In both CSCT and CCTA, in-plane resolution was 0.4-0.5 $\times$ 0.4-0.5 mm.

The set of 250 cardiac CT exams was divided into two sets. The first 50 exams were used to train an algorithm that detects bounding boxes around the heart. The remaining 200 exams were used to train and evaluate a voxel classification algorithm that identifies CAC in these bounding boxes. Two expert observers provided annotations in all (observer O_1) or in a subset (observer O_2) of the exams.

In each of the 50 cardiac CT exams used to train the bounding box detection algorithm, observer O_1 manually determined a 3D rectangular bounding box around the heart in the CCTA scan. This bounding box included the pericardial sac, from below the pulmonary artery to the apex in the craniocaudal direction.

In each of the 200 cardiac CT exams used to train and evaluate the voxel classification algorithm, manual reference annotations for CAC were obtained in the CCTA and the CSCT scan. Manual annotations in the CCTA scans were obtained similarly to the methods proposed in [20, 19]. The expert observer first manually identified a point in the center of the ascending aorta at the level of the origin of the left coronary artery. This point was automatically grown to a 200 mm^3 volume of interest (VOI). The mean ($mean_{aorta}$) and standard deviation (SD_{aorta}) of HU values in this ROI were used to compute a patient-specific threshold $mean_{aorta} + 3SD_{aorta}$. The expert observer then marked calcification in the coronary artery by a mouse click on a single voxel, which was followed by automatic 3D region growing of voxels with density above the defined threshold. In addition, to compare obtained CAC scores in CCTA to the clinically used standard, CAC in CSCT scans was also manually identified with a clinically used threshold of 130 HU [9].

Observer O_1 annotated CAC in all 200 cardiac CT exams. These annotations were considered the reference standard, used for training and evaluation of the voxel classification algorithm. Observer O_2 annotated CAC in a subset of 100 cardiac exams. These annotations were used to determine inter-observer variability and for comparison with the automatic method.



(a) (b)

Figure 5.2: CAC-like voxels in CCTA: (a) aortic valve calcification (solid white arrows), calcification in the descending aorta (dashed white arrow), and (b) CAC in the left anterior descending artery (solid white arrows).

5.3 Method

CAC was identified by voxel classification. Besides CAC, a typical CCTA scan contains many other voxels of appearance similar to CAC. These include extracardiac lesions like bones such as ribs, calcifications in the descending aorta and calcified lymph nodes, as well as intracardiac calcifications such as those in the mitral and aortic valve (Fig. 5.2). In addition, coronary artery lumen is often highly attenuated, hence resembling CAC.

The proposed algorithm is illustrated in Fig. 5.3. First, a bounding box around the heart is determined. This excludes most extracardiac calcifications and allows further analysis within this VOI only. Next, voxels in the VOI are classified with a pair of CNNs (ConvPair), which share the same structure but have differently trained parameters. The first CNN (CNN_1) detects voxels likely to be CAC among all candidate voxels. The second CNN (CNN_2) separates CAC from CAC-like voxels, such as attenuated coronary artery lumen and aortic calcifications. Finally, identified CAC is quantified.

5.3.1 Preprocessing

CCTA scans are generally acquired with a standardized scan length in the craniocaudal direction ranging from mid-pulmonary artery to diaphragm [107]. However, their field of view in the transverse plane is less standardized. Some scans might contain the ribs and spine, while others may be closely cropped around the heart.

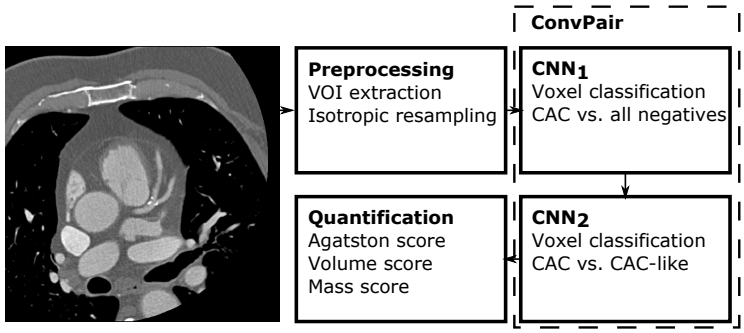


Figure 5.3: Overview of the proposed system. The image is preprocessed by determination of a bounding box around the heart and isotropic resampling. Voxels resembling CAC are extracted and subsequently classified by a pair of CNNs (ConvPair). CNN_1 identifies CAC-like voxels, among which CNN_2 distinguishes between CAC and other CAC-like candidates. The resulting segmentation is quantified using the CAC Agatston, volume and mass score.

To reduce this variation, and to allow analysis only within the VOI, the field of view is standardized by finding a 3D rectangular bounding box around the heart. This bounding box is automatically determined using our previously developed algorithm described by De Vos et al. [103]. The algorithm uses three independent CNNs, each determining the presence of the heart in the axial, sagittal or coronal view. A rectangular 3D bounding box around the heart is obtained by combining posterior probabilities obtained for axial, sagittal and coronal image slices.

CNNs label a target voxel based on a square or cubic input patch centered at that voxel. For this, it is beneficial to have identical receptive fields along all image axes. Therefore, images cropped to the determined bounding box are resampled with B-spline interpolation to 0.45 mm isotropic voxels – the standard slice spacing in our data set. Finally, to allow robust CNN training, all data is rescaled to the mean and standard deviation of HU values in the training images.

5.3.2 Voxel classification

All voxels in the VOI are considered candidates for CAC. The proposed CNNs take one or multiple patches of size w voxels centered at the candidate voxel as input and extract features based on that input. Using these features, the voxel is assigned a probability p_{CAC} of being CAC. Input patches are either 2.5D, i.e. three 2D patches from orthogonal image planes centered at the voxel, or 3D, i.e. volumetric patches centered at the voxel.

Features in a CNN are typically extracted by a stack of convolution layers, while classification is done by a stack of fully-connected hidden layers. The proposed method uses a purely convolutional CNN architecture for both the feature extraction stack and the classification stack, suitable for both 2.5D and 3D input. Fig. 5.4 shows an overview of the proposed architecture.

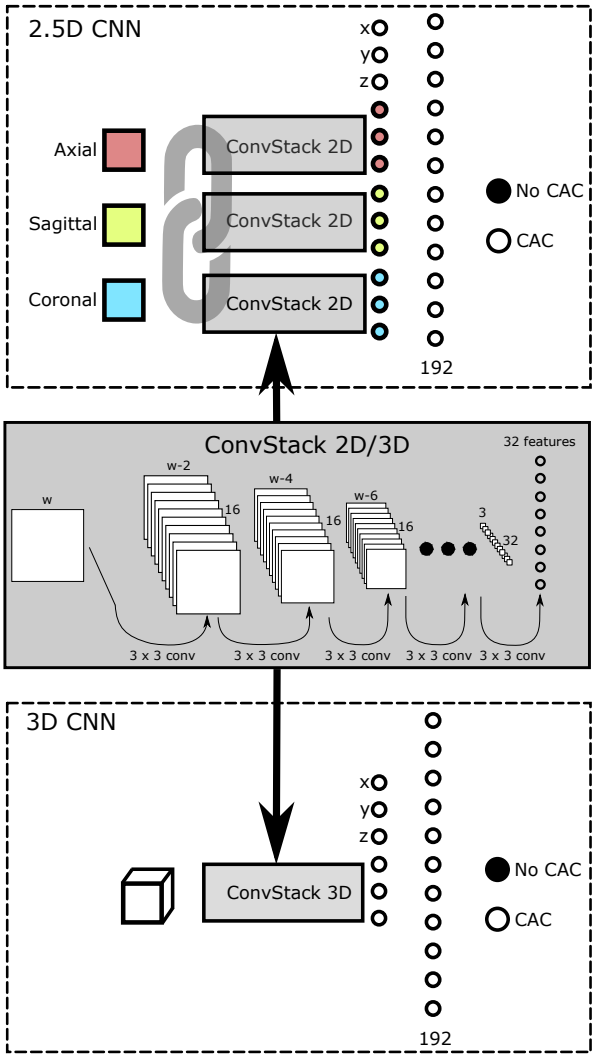


Figure 5.4: Proposed convolutional neural network (CNN) architecture for 2.5D and 3D input. The network consists of a stack of feature extraction layers and a stack of classification layers. The feature extraction stack (ConvStack) has the same hyperparameters in both architectures. It consists of a sequence of layers with 3×3 voxel (in 2D) or $3 \times 3 \times 3$ voxel (in 3D) convolution kernels. The 2.5D CNN combines features from three identical 2D ConvStacks with shared weights, each processing an input patch from a different orthogonal viewing direction, i.e. axial, sagittal and coronal. These input patches are centered at the target voxel. The 3D CNN uses volumetric features extracted from a 3D input patch centered at the target voxel. In both the 2.5D and the 3D CNN, features are concatenated with x, y and z location features and connected to an output layer through one hidden layer.

Depending on the size and the dimensionality of the input patches, the architecture of the feature extraction stack is generated as follows. Each convolutional layer with the exception of the last one consists of 16 small convolution kernels of 3×3 voxels in 2.5D or $3 \times 3 \times 3$ voxels in 3D. Choosing multiple stacked small convolution kernels over one larger convolution kernel has been shown to have two advantages [108]. First, more stacked layers contain more non-linear activation layers, hence allowing the network to become more discriminative. Second, stacking small kernels reduces the number of trainable parameters, and hence the chance of over-fitting.

Convolutions are valid, i.e. no zero-padding is applied after convolution, so that each convolution reduces the input size by 2 voxels along each axis. In the final convolution layer, 32 convolution kernels reduce the input size to 1 voxel along each axis. The 32 obtained features are used for classification. Each convolution layer was followed by a rectified linear unit (ReLU) activation function [109].

The convolutional stack does not include any max-pooling downsampling layers. These layers are typically used in image classification and object detection to rapidly decrease the input size, to reduce the number of weights in the network to prevent overfitting and to introduce spatial invariance. However, the spatial invariance introduced by pooling could mean that neighboring voxels are assigned the class label that is most expressed in that location. This in turn could lead to over- or undersegmentation of CAC lesions, which are generally small. In addition, the absence of pooling layers means that the convolutional stack is purely convolutional. Hence, the convolutions may be applied to full images, thereby avoiding redundant convolution operations.

A 2.5D CNN contains three convolutional stacks, which independently process axial, sagittal and coronal input. The three networks share weights, i.e. one 2D network was used for feature extraction in the three orthogonal planes, similar to the shared weight multi-scale approach in [110]. Tying the weights in these networks reduces the number of features and allows robust generic texture feature extraction. A 3D CNN contains one volumetric convolutional stack.

The features extracted by the convolutional stack are used to classify the target voxel as either CAC or non-CAC. The extracted features might only provide limited spatial information. However, studies on automatic CAC scoring in non-contrast-enhanced cardiac CT have shown that location information is essential for CAC identification [14, 13, 15]. Therefore, a normalized heart coordinate system is used to describe the location of each voxel within the VOI. In this coordinate system, the origin is located at the center of the VOI and -1 and 1 are positioned at the boundaries of the VOI along each axis. For each candidate, the x-, y- and z-coordinate are determined as location features. These location features are concatenated to the texture features derived by the network to provide a feature vector.

The feature vector serves as input to one fully-connected hidden layer. This hidden layer is connected to an output layer with a softmax function to predict probabilities p_{CAC} and $1 - p_{CAC}$. To regularize the network, Dropout is applied before and

after the hidden layer.

5.3.3 Training strategy

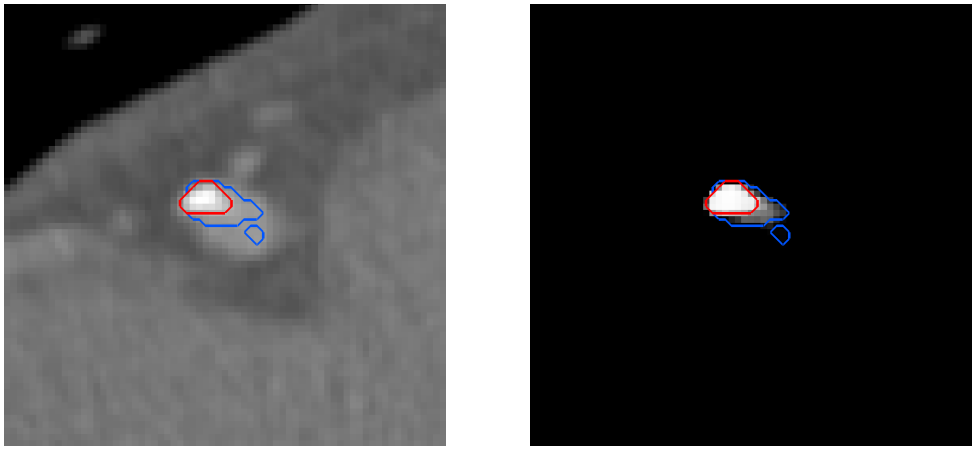
CCTA scans contain many more negative (background) than positive (CAC) samples, posing a heavily unbalanced classification problem. Identifying CAC among all voxels in a cardiac VOI in CCTA poses two challenges. The vast majority of negatives such as those representing lung or fatty tissue, share very few similarities with CAC. Hence, given sufficiently descriptive features, they might easily be discarded. Other negatives, such as bone (e.g. sternum), calcifications in the ascending aorta and coronary artery lumen enhanced by contrast material, are more challenging to distinguish from CAC. Our method uses a ConvPair, a pair of CNNs, each of which have a specific task. CNN_1 focuses on detection of CAC-like voxels, CNN_2 identifies CAC voxels among these candidates.

The two CNNs are trained in sequence. CNN_1 is trained first, using all voxels in the VOIs of the calcium scoring training images. This CNN learns to discard the vast majority of negative voxels. For each calcium scoring training image, a CAC candidate mask is obtained using CNN_1. This mask contains CAC-like voxels, but no negatives such as lung tissue or fatty tissue. Subsequently, CNN_2 is trained using only the samples in the CAC candidate mask. To leverage already learned knowledge, CNN_2 is initialized as the final version of CNN_1 and training is resumed using the samples from the CAC candidate mask. Hence, CNN_1 and CNN_2 share their architecture but not their trainable parameters.

During testing, CNN_1 and CNN_2 in a ConvPair can be evaluated sequentially, i.e. by first obtaining a candidate mask from CNN_1 and classifying only those candidate voxels in the mask with CNN_2. Alternatively, both CNNs may be merged into one network with two parallel stacks of layers, where the first stack contains CNN_1 and the second stack contains CNN_2. To obtain a probabilistic CAC map, p_{CAC} values generated by CNN_1 are thresholded and the resulting binary image is multiplied with the p_{CAC} values generated by CNN_2.

5.3.4 Implementation

All CNNs were implemented using Theano [111]. The purely convolutional nature of the networks was exploited to rapidly process an image of size $nx \times ny \times nz$ voxels. In 2.5D, full image slices along the three principal axis were independently processed. This resulted in three 4D texture feature tensors of size $nx \times ny \times nz \times nf$, where $nf = 32$ in the proposed architecture (Fig. 5.4). These tensors were concatenated along the last dimension. Similarly, in 3D one subimage of size $nx \times ny \times w$ voxels was processed for every axial slice, to obtain a texture feature tensor. In both 2.5D and 3D, the texture feature tensor was concatenated with a $nx \times ny \times nz \times 3$ location feature tensor. Fully connected hidden layers were implemented as convo-



(a) CCTA scan **(b)** p_{CAC}

Figure 5.5: (a) CCTA scan showing a right coronary artery (RCA) CAC lesion, and (b) aligned posterior CAC probability p_{CAC} map. The blue contour shows manual lesion segmentation based on a patient-specific intensity threshold in the CCTA scan, and the red contour shows automatic lesion segmentation based on a threshold θ_{CAC} in the posterior probability map.

lutions with kernel size 1. Hence, two consecutive convolutions for the hidden and output layer resulted in a probability distribution for each voxel.

5.3.5 Evaluation

Reference lesions were segmented using 26-connected 3D region growing of voxels above a patient-specific intensity threshold [20]. Although this patient-specific threshold identifies CAC better than a global threshold (Fig. 5.1), it is based on aortic attenuation, which can differ from coronary attenuation due to e.g. luminal narrowing, imaging artifacts or partial volume effects. Hence, intensity-based region growing may under- or oversegment CAC lesions (Fig. 5.5). Therefore, automatically obtained lesions were not segmented based on attenuation, but based on posterior probabilities, using 26-connected 3D region growing of voxels with a CAC probability $p_{CAC} \geq \theta_{CAC}$, where p_{CAC} was predicted by the ConvPair and θ_{CAC} was determined using ROC analysis. Lesions smaller than 1.0 mm^3 were discarded as these likely represented noise.

Given that CAC lesions in CCTA were created differently in the reference and automatic results, they might not contain the same voxels. Therefore, true positive lesions were automatically found lesions having overlap with the reference lesions. False positive lesions were those in the automatic result having no overlap with the reference lesions. False negative lesions were lesions identified in the reference that had no overlap with any automatically found lesion.

In addition to an evaluation of the ability of the method to detect individual lesions, its ability to determine per patient Agatston, volume and mass CAC scores was established. The Agatston score weighs calcified plaque area by peak intensity, and was computed as $\sum_{s \in S} a_s \cdot d_s \cdot \frac{\Delta_z}{3.0}$, where S is the set of slices containing the lesion, a_s is the lesion area (in mm^2) in s, d_s is a density factor based on the lesion's peak intensity in s (130–199 HU: 1, 200–299 HU: 2, 300–399 HU: 3, $\geq$ 400 HU: 4), and Δ_z is the image slice increment. A linear scaling factor $\frac{\Delta_z}{3.0}$ is standardly used to correct for image slice increments different than the 3.0 mm increment for which the method was originally developed [9]. The CAC volume score (in mm^3) quantifies CAC volume and was computed as the number of identified voxels multiplied by voxel volume (in mm^3). The CAC equivalent mass score weighs lesion attenuation linearly and was computed as the product of the volume of a lesion and its mean intensity [11]. Per lesion CAC scores were summed to obtain per patient scores.

Automatically obtained volume and mass scores in CCTA were compared with reference volume and mass scores in CCTA by observer O_1. In current clinical practice, CAC burden is determined in CSCT. Hence, automatically obtained volume and mass scores in CCTA were compared with reference volume and mass scores in CSCT. In addition, manual volume and mass scores defined by observer O_1 and observer O_2 in CCTA scans were compared with the reference scores in CSCT. Finally, to establish interobserver agreement, volume and mass scores in CSCT and CCTA scans were compared between the observers. Agreement between two measurements was determined using the intra-class correlation coefficient (ICC) for absolute agreement, with 95% confidence interval. In addition, Bland-Altman plots were generated and the bias and limits of agreement (±1.96 SD) were reported.

In clinical practice, patients are assigned to a CVD risk category based on their Agatston score. Because CVD risk categorization is not defined for Agatston scores in CCTA, agreement between standard Agatston score based risk categorization in CSCT and the Agatston score in CCTA was determined as follows. Patients were first categorized based on their CSCT Agatston score according to standard risk categories (Very low: 0, Low: 1-100, Intermediate: 101-400, High: >400). Thereafter, patients were ranked based on their CCTA Agatston score. Based on this ranking, to each of four CCTA risk categories the same number of patients was assigned as in the corresponding CSCT category. Patients who ended up in the same category based on both CSCT and CCTA Agatston score were correctly categorized, patients who were assigned to a different category were incorrectly categorized. Categorization accuracy and Cohen's linearly weighted κ were computed.

5.4 Experiments and results

The set of 250 exams was divided into four sets. First, a set of 50 exams was used to train the bounding box extraction algorithm. Second, a training set of 90 exams was used to train the CNN pairs for CAC classification. Third, a validation set of 10

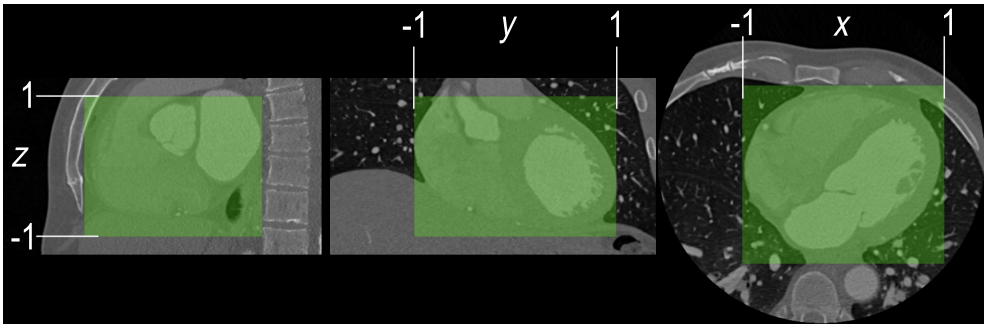


Figure 5.6: Automatically obtained heart bounding box in CCTA in sagittal, coronal and axial view, respectively. The bounding box reduces the volume of interest (VOI) by up to 80%. In addition, the boundaries provide a normalized coordinate space with the origin in the center of the box.

Table 5.1: Details of the evaluated CNN pairs. Number of layers, without input and output layers (#Layers), number of trainable weights (#Weights), average and SD processing time per image for CNN_1 in s ($Time_1$) and for CNN_2 in s ($Time_2$).

		#Layers	#Weights	$Time_1$	$Time_2$
$w = 15$	2.5D	8	35,986	42 ± 5	26 ± 4
	3D	8	56,242	52 ± 6	107 ± 44
$w = 25$	2.5D	13	47,586	46 ± 7	28 ± 4
	3D	13	90,882	147 ± 15	201 ± 113

exams was used to optimize hyperparameters of the CNNs. Finally, a test set of 100 exams was only used to evaluate the performance of the method. For the test set, annotations by both observer O_1 and O_2 were available.

5.4.1 Bounding box extraction

Bounding box extraction took 6.9 ± 0.5 s, discarding up to 80% of the original CCTA image. Fig. 5.6 shows an example of an automatically determined bounding box. Each VOI contained around 20,000,000 negative voxels and on average 800 positive voxels. Visual inspection of the results showed that in all cases the bounding box contained the whole heart.

5.4.2 Experimental settings

Network weights were initialized according to the procedure specified by [112]. Dropout probability in fully connected layers was set to $p = 0.5$ [113]. The categorical cross-entropy between reference and predicted labels was minimized using stochastic gradient descend with Nesterov momentum and learning rate $\alpha = 0.001$.

Table 5.2: Lesion identification with respect to the reference standard in CCTA. Four ConvPairs with dimensionality 2.5D or 3D, and input patch size $w = 15$ or $w = 25$ were trained. Ensembles of ConvPairs were formed by averaging of probabilities. Bullet points indicates membership of an ensemble. Lesion identification sensitivity is reported, as well as the average number of false positive (FP) lesions per scan.

2.5D		3D			
$w = 15$	$w = 25$	$w = 15$	$w = 25$	Sens.	FP/scan
•				68%	0.90
	•			72%	0.85
		•		67%	1.69
			•	72%	1.21
•	•			71%	0.64
		•	•	69%	0.77
•		•		71%	0.68
	•		•	71%	0.57
•	•	•	•	71%	0.48

CNN_1 was trained with 200,000 mini-batches, CNN_2 with 100,000 mini-batches. Mini-batches were balanced, containing 64 negative and 64 positive samples. Probability maps generated by CNN_1 were thresholded at $p_{CAC} \geq 0.5$ to provide a mask for CAC detection. All models were trained and tested on single NVIDIA Titan X GPUs.

Four ConvPairs were trained, namely for 2.5D and 3D input patches with size $w = 15$ and $w = 25$. Table 5.1 lists the number of layers and the number of trainable parameters of each ConvPair architecture, as well as the average time required for processing by its components CNN_1 and CNN_2. For 2.5D input, the networks with input size $w = 25$ took slightly more time than the network with size $w = 15$. In 3D, the difference in required processing time between input sized $w = 15$ and $w = 25$ was larger.

5.4.3 Lesion identification

ConvPairs are specified by input dimensionality and input size, and results are presented for merged output of their members CNN_1 and CNN_2. Automatically extracted lesions were compared to the reference standard in CCTA test scans, which in total contained 260 CAC lesions.

Table 5.2 lists sensitivity for lesion identification, as well as the average number of false positive (FP) errors per scan. Results are shown for individual ConvPairs, as well as ensembles combining ConvPairs with different input sizes and dimensionality. Recent large-scale 2D natural image classification challenges have been dominated by ensembles of CNNs. It has been shown that combinations of CNNs

improve classification results, particularly when CNNs have been trained with different hyperparameters.

The results indicate that among architectures with the same dimensionality, those with an input size $w = 25$ perform better than those with an input size $w = 15$. Among architectures with the same input size, 2.5D networks obtain better results than 3D network. The strongest individual architecture combines the largest input size $w = 25$ with a 2.5D input representation.

In all cases, ensembles make fewer FP errors per scan at a similar sensitivity level. The ensemble with $w = 25$ outperforms the ensemble with $w = 15$, and the 2.5D ensemble outperforms the 3D ensemble. An ensemble of all ConvPairs provides the best result, at 71% lesion sensitivity and 0.48 FP error per scan.

Fig. 5.7 shows typical examples of false negative (FN) and FP errors. A CAC lesion of low density compared to the surrounding lumen in the left anterior descending (LAD) artery was missed (Fig. 5.7a), as well as a CAC lesion that was deformed by a motion artifact in the RCA (Fig. 5.7b). In a small number of scans, all ConvPairs and ensembles of ConvPairs made the same errors. One scan contained a stent in the RCA, parts of which were incorrectly identified as CAC (Fig. 5.7c). A second scan contained large calcified lymph nodes, parts of which were incorrectly identified as CAC (Fig. 5.7d). Several other false positive errors were due to high HU values in the coronary artery lumen.

The effect of CNN_1 and CNN_2 on voxel classification was investigated. Fig. 5.8 shows typical probabilistic maps obtained by CNN_1 and CNN_2, obtained with 2.5D input with size $w = 25$. While CNN_1 detected CAC, it also detected lumen, bone in the rib and aortic calcification (Fig. 5.8b). CNN_2 correctly identified CAC, but also gave a weak response in voxels dissimilar to those which were used to train CNN_2, i.e. high-density voxels similar to CAC (Fig. 5.8c). Because the network was fine-tuned only with samples similar to CAC, it had lost its ability to classify other candidates. Finally, in the merged result of CNN_1 and CNN_2 (Fig. 5.8d), only CAC received a high probability.

To investigate the effect of the normalized x, y, z–coordinates on CAC identification, an additional ConvPair was trained with the best performing architecture, i.e. 2.5D input with $w = 25$, adapted to omit the feature coordinates. ROC analysis showed that at all sensitivity levels, this ConvPair made slightly more FP errors than the ConvPair with location features.

5.4.4 Per patient CAC quantification

For each patient, the volume and mass scores in CCTA were determined by observer O_1, observer O_2 and the automatic method. Also, for the automatic method, Agatston scores in CCTA were determined. Automatic results were based on the complete ensemble of trained architectures.

Fig. 5.9 shows a Bland-Altman plot for the agreement between reference CAC scores in CCTA and automatically obtained CAC scores in CCTA. Automatically ob-

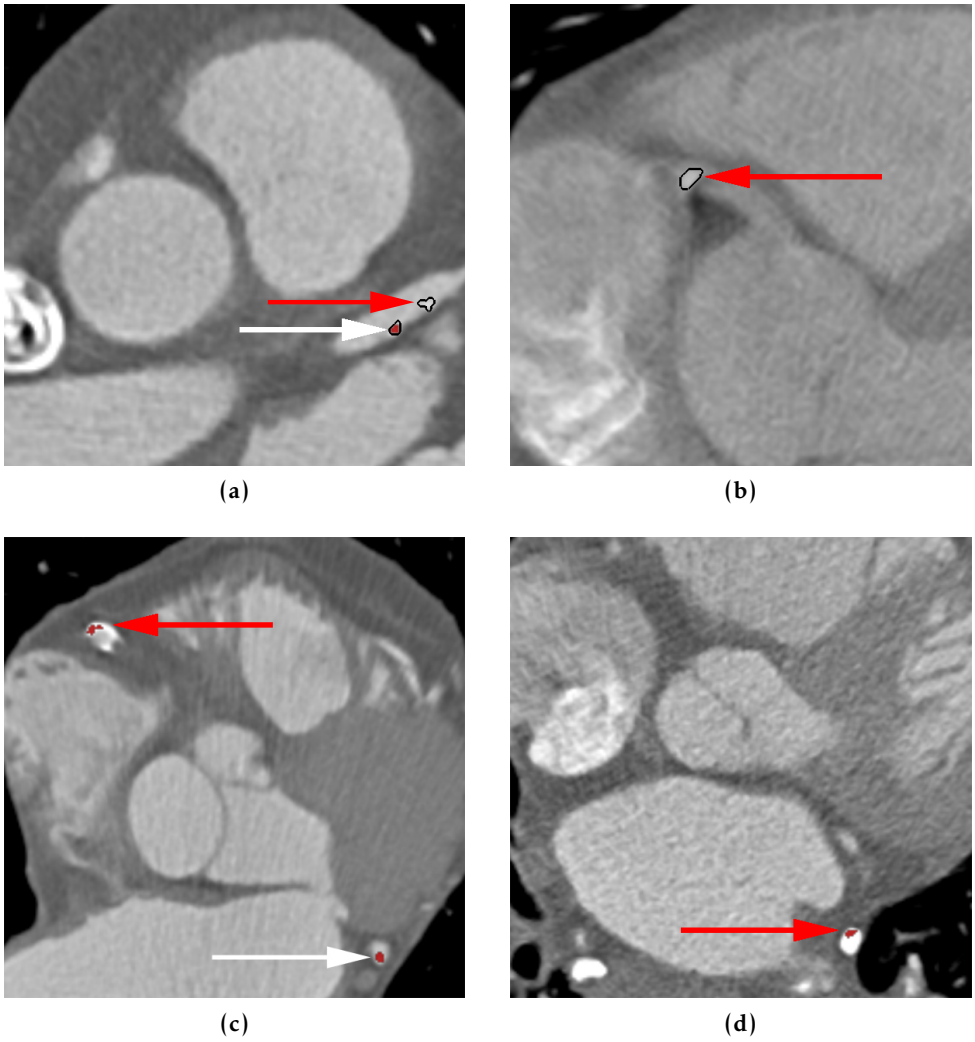
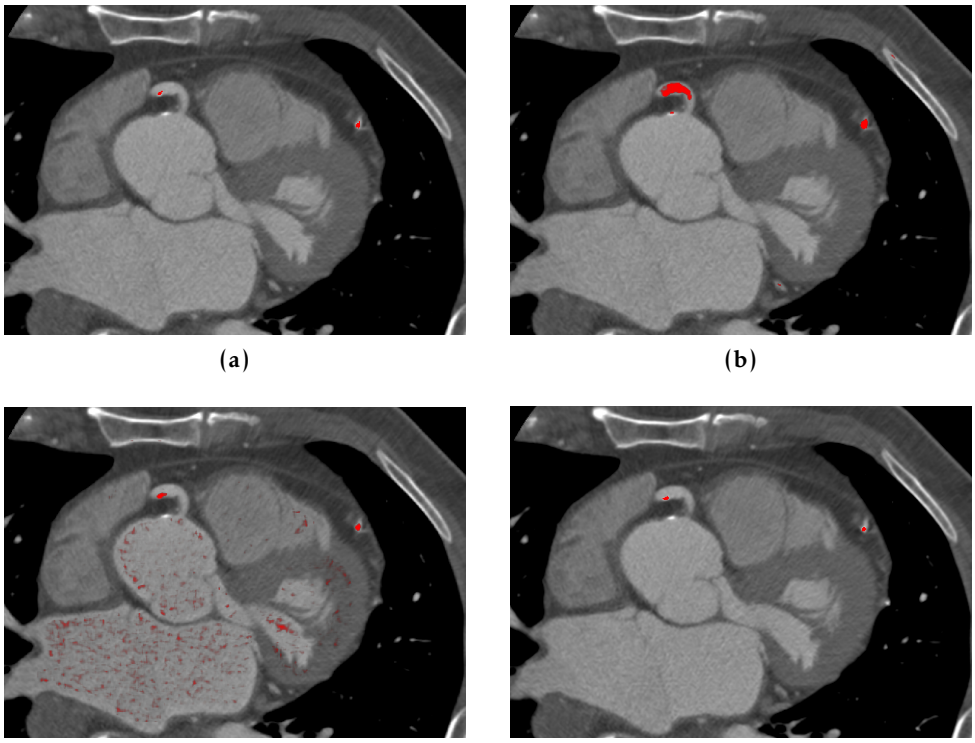


Figure 5.7: Examples of false negative (FN) and false positive (FP) errors. (a) Two CAC lesions in the left anterior descending artery. One was missed by the algorithm (red arrow), one was found (white arrow). (b) CAC lesion affected by motion artifact in right coronary artery (RCA) which was not identified by the algorithm. (c) CAC lesion in the left circumflex artery which was correctly identified (white arrow), and part of a stent which was incorrectly identified as CAC (red arrow). (d) Calcified lymph node, part of which was incorrectly identified as CAC (red arrow).



(a) **(b)**

(c) **(d)**

Figure 5.8: (a) Overlay showing reference annotation with CAC lesions in the left anterior descending (LAD) and right coronary artery (RCA) in bright red. (b) Probabilistic map generated by CNN_1. The map contains high probabilities (bright red) for CAC, but also for coronary lumen and for calcification in the ascending aorta. (c) Probabilistic map generated by CNN_2. The map shows high probabilities for CAC (bright red), while the probabilities for coronary lumen and aortic calcification are zero. However, CNN_2 also assigns CAC probabilities to blood in the left atrium, left ventricle and ascending aorta (dark red), as it was specifically trained on CAC-like voxels. (d) thresholded merged probabilistic maps of CNN_1 and CNN_2, showing identified CAC voxels (bright red).

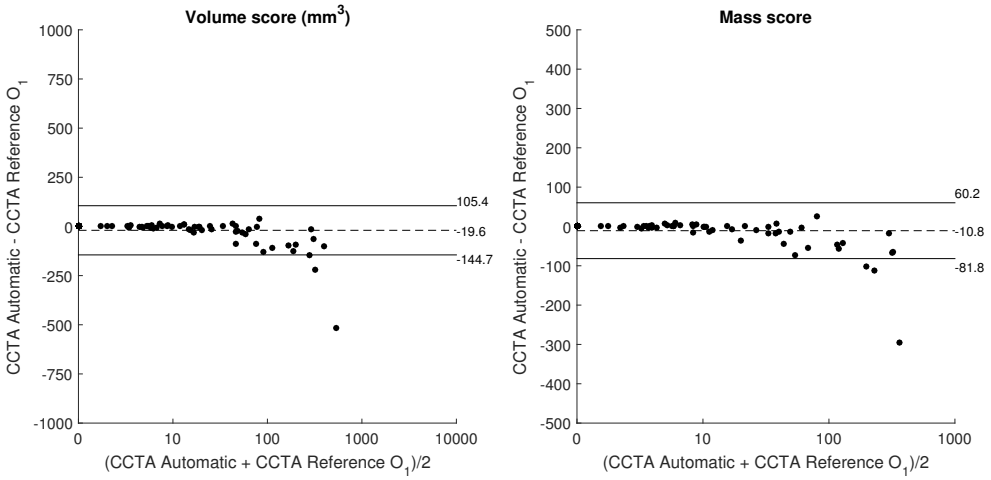


Figure 5.9: Bland-Altman plots comparing automatically obtained CAC volume and mass scores in CCTA with reference annotations by observer O_1 in CCTA. Bland-Altman bias and limits of agreement are indicated.

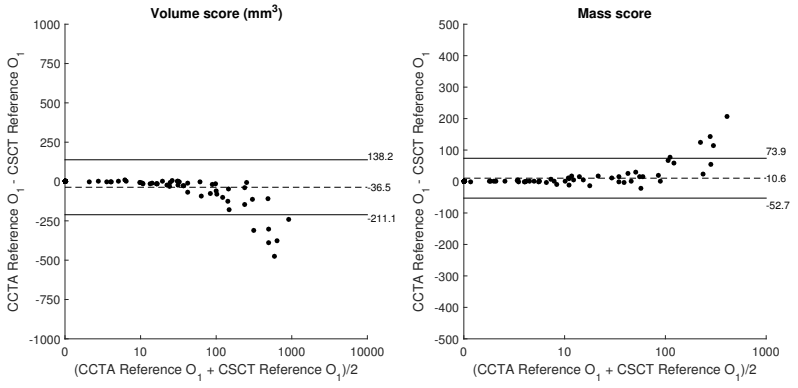
tained CAC scores were typically lower than those in the reference standard. Bland-Altman bias and limits of agreement were -19.6 (-144.7–105.4) mm^3 for CAC volume and -10.8 (-81.8–60.2) for CAC mass score. The ICC was 0.768 (0.660–0.842) for CAC volume and 0.872 (0.808–0.915) for CAC mass score.

Fig. 5.10 shows Bland-Altman plots for the agreement between reference CAC scores in CSCT and CAC scores in CCTA determined by observer O_1, observer O_2, and the automatic method. One patient was left out of the statistical analysis due to a large motion artifact in CSCT.

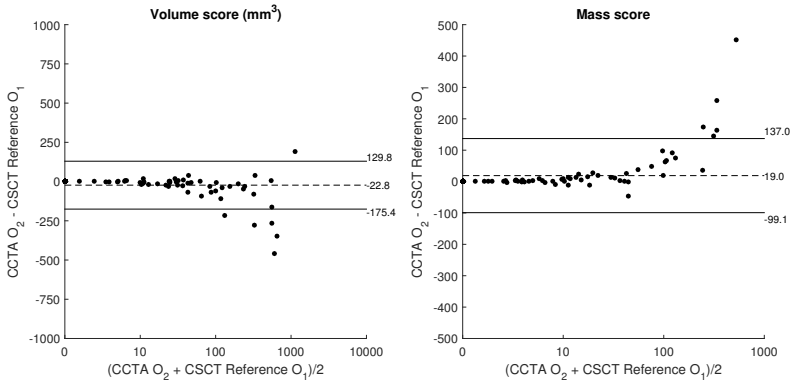
CAC volume in CCTA was lower than the reference CAC volume in CSCT. This effect was stronger for the automatic method than for the observers. Bland-Altman bias and limits of agreement were -36.5 (-211.1–138.2) mm^3 and -22.8 (-175.4–129.8) mm^3 for observers O_1 and O_2, and -56.1 (-319.4–207.2) mm^3 for the automatic method. The ICC was 0.828 (0.719–0.891) and 0.900 (0.848–0.934) for observers O_1 and O_2, and 0.538 (0.347–0.679) for the automatic method.

For CAC mass score, values in CCTA were lower than the reference in CSCT for the two observers, but not for the automatic method. Bland-Altman bias and limits of agreement were -10.6 (-52.7–73.9) for observer O_1, 19.0 (-99.1–137.0) for observer O_2, and -0.2 (-38.7–38.3) for the automatic method. The ICC was 0.895 (0.837–0.932) for observer O_1, 0.761 (0.650–0.838) for observer O_2, and 0.944 (-0.918–0.962) for the automatic method.

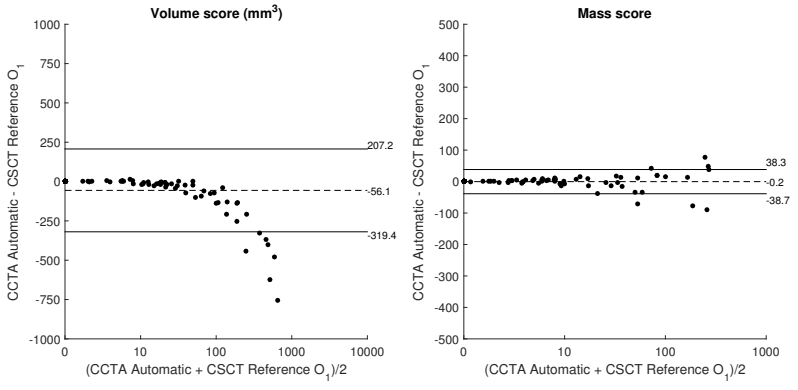
The confusion matrix in Table 5.3 compares Agatston-score based risk categorization in the reference CSCT annotations and CVD risk categorization based on automatically determined Agatston scores in CCTA. CVD risk categorization accuracy was 84%, with linearly weighted $\kappa = 0.83$. No patient was more than one cat-



(a) Observer O_1



(b) Observer O_2



(c) Automatic method

Figure 5.10: Bland-Altman plots comparing reference CAC volume and mass scores in CSCT with annotations in CCTA by (a) observer O_1, (b) observer O_2 and (c) the automatic method. Bland-Altman bias and limits of agreement are indicated.

Table 5.3: Agreement in cardiovascular risk categorization based on Agatston score categories in the CSCT reference standard (I: 0, II: 1-100, III: 101-400, IV: >400) and derived categories in the automatically obtained CCTA Agatston score.

		Reference				
		I	II	III	IV	Total
Automatic	I	**43**	4	0	0	47
	II	5	**23**	3	0	31
	III	0	3	**10**	1	14
	IV	0	0	1	**7**	8
	Total	48	30	14	8	100

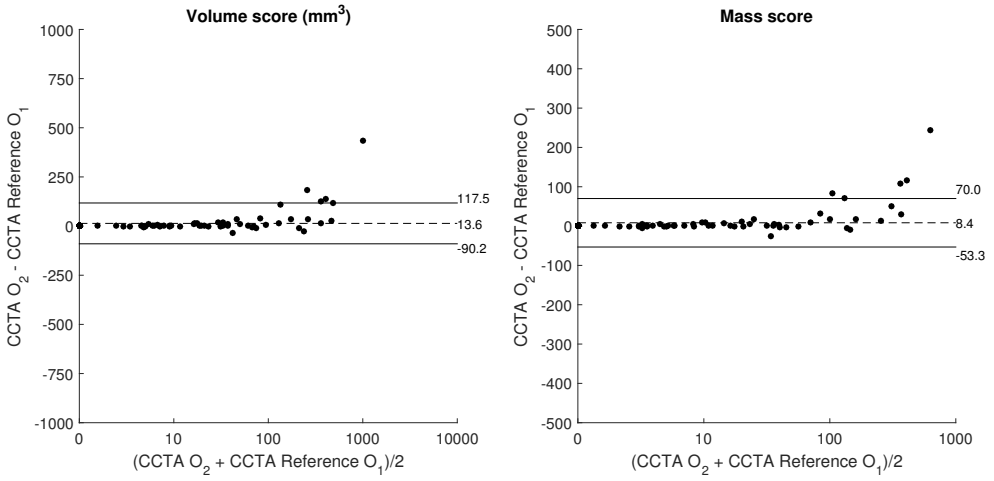
egory off. The test set contained 48 patients with zero CAC and 52 patients with a positive CAC score. The automatic method identified 43/48 patients with zero CAC. Conversely, the automatic method correctly identified 48/52 patients with a positive CAC score. Missed CAC lesions in the remaining patients were small and low-density lesions, with the exception of one patient whose scan contained a CAC lesion in the RCA which was deformed due to a motion artifact and therefore missed by the algorithm (Fig. 5.7b).

5.4.5 Interobserver agreement

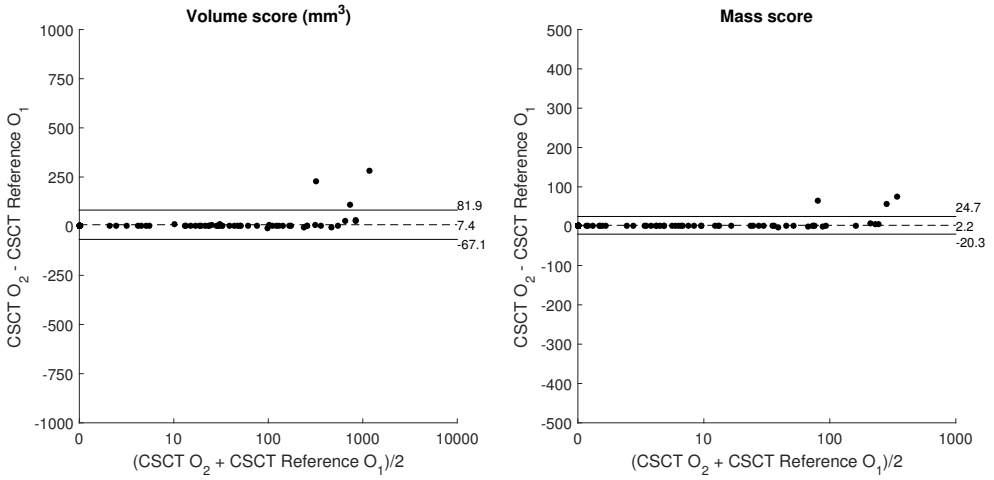
Both observers scored CAC in the full test set. To score CAC in CCTA, a patient-specific CAC threshold was determined using a manually placed ROI in the ascending aorta. This threshold showed excellent agreement between the two observers (ICC 0.994 [0.990–0.997]). The thresholds ranged from 332 to 898 HU for observer O_1 and from 333 to 930 HU for observer O_2.

Agreement between CAC scores of the two observers in both CCTA and CSCT was excellent. For CAC volume and mass in CCTA, the ICC between the two observers was 0.928 (0.891–0.952) and 0.949 (0.922–0.966), respectively. Bland-Altman analysis (Fig. 5.11a) had bias and limits of agreement 13.6 (-90.2 – 117.5) mm^3 and 8.4 (-53.3–70.0) for CAC volume and CAC mass, respectively. For CAC volume and mass in CSCT, the ICC between the two observers was 0.983 (0.974–0.988) and 0.982 (0.974–0.988). Bland-Altman analysis (Fig. 5.11b) had bias and limits of agreement -7.4 (-67.1–81.9) mm^3 and 2.2 (-20.3–24.7) for CAC volume and CAC mass, respectively.

Observer O_2 identified 45/48 patients with zero CAC score in CCTA as annotated by observer O_1, the reference. Conversely, observer O_2 identified 51/52 patients with a positive CAC score as annotated by observer O_1.



(a) Inter-observer agreement in CCTA



(b) Inter-observer agreement in CSCT

Figure 5.11: Bland-Altman plots comparing (a) CAC volume and mass scores in CCTA and (b) CAC volume and mass scores in CSCT between observer O_1 and observer O_2. Bland-Altman bias and limits of agreement are indicated.

Table 5.4: Previously published results on automatic CAC scoring in CCTA. For each study, the following are reported: the number of scans in the test set (#Scans), the evaluated CAC quantification score (Score, AS = Agatston score [modified for CCTA], MS = mass score), Pearson correlation (Pearson ρ), intra-class correlation coefficient (ICC), Bland-Altman mean and limits of agreement in units or percentages (Bland-Altman), CVD risk categorization accuracy and weighted Cohen's κ (CVD risk), lesion identification sensitivity (Sens.) and false positive errors per scan (FP/scan) are listed.

	#Scans	Score	Pearson ρ	ICC	Bland-Altman	CVD risk	Sens. FP/scan
Schuhbaeck et al. [68]	44	AS_{CCTA} vs. AS_{CSCT}	0.94	–	-56 (-518–407)	88.6% κ = 0.87	–
Ahmed et al. [63]	100	AS_{CCTA} vs. AS_{CSCT}	0.949	0.863	-3 (-174 – 168) %	76.0% κ = 0.588	–
Eilot et al. [64]	263	AS_{CCTA} vs. AS_{CSCT}	0.95/0.91	–	-1 (-80 – 78) %	82.7%	0.94, 0.9
Teßmann et al. [102]	53	AS_{CCTA} vs. AS_{CCTA}	0.95	–	–	–	0.94, 0.9
Mittal et al. [65]	165	–	–	–	–	–	0.70, 0.1
Wesarg et al. [67]	10	–	–	–	–	–	1.00, –
Proposed method	100	MS_{CCTA} vs. MS_{CSCT}	0.950	0.944	-0.2 (-38.7–38.3)	84% κ = 0.83	0.72, 0.48

5.4.6 Comparison with previous methods

The proposed method was compared with results reported by other previously published algorithms on CAC scoring in CCTA. Table 5.4 lists, where available, the number of scans used for evaluation in each study, the evaluation criterion, Pearson's ρ correlation, ICC and/or Bland-Altman statistics between manual and reference scores, CVD risk categorization accuracy and lesion detection sensitivity, as well as the average number of FP errors per image. The listed results cannot be directly compared, as different data sets, different CAC quantification metrics, different correlation metrics and different CVD risk categories were used. E.g. some studies specifically excluded patients with moderate or poor image quality, patients without CAC or patients with stents. All methods except the proposed method require coronary artery extraction for CAC detection.

5.5 Discussion

A method for automatic coronary artery calcium scoring in coronary CT angiography employing convolutional neural networks has been presented. In contrast to previously proposed methods for CAC scoring in CCTA, our method does not require coronary artery extraction. Instead, CAC voxels are directly identified using pairs of CNNs.

Automatically obtained as well as reference CAC volume scores in CCTA were lower than in CSCT. This is in accordance with previous studies [114, 20]. However,

automatically obtained CAC mass scores showed a strong correlation (ICC 0.944 [-0.918–0.962]) with reference CAC mass scores in CSCT. A comparison of CVD risk categories based on reference Agatston scores in CSCT and automatically obtained Agatston scores in CCTA showed excellent agreement, with 84% of patients assigned to their reference risk category (κ = 0.83). Hence, patients with a high reference Agatston score in CSCT also had a high automatically determined Agatston score in CCTA. In addition, discrimination between patients with zero CAC and patients with a positive CAC score was good. Large scale studies have shown that patients with zero CAC have an excellent prognosis [80], underlining the clinical relevance of this distinction.

False positive errors were caused by high intensity voxels in the coronary artery lumen. Lumen attenuation may be very different among patients, thereby posing a challenge for accurate CAC segmentation. In our test set, CAC detection thresholds determined based on the attenuation of the ascending aorta ranged from 333 to 930 HU. The method made some false positive errors, e.g. calcified lymph nodes, that are less likely to occur in methods using coronary artery extraction. Nevertheless, although our method does not use coronary artery extraction, it is likely that the CNN implicitly learns a representation of tubular structures. Calcifications in the coronary arteries were identified, but calcifications in the aorta, with similar intensity and shape characteristics but a different context, were not a source of false positive errors.

A VOI containing the heart was determined using a CNN-based bounding box extraction algorithm. Although this VOI primarily contained cardiac structures, non-cardiac structures such as ribs (see Fig. 5.6) were occasionally partially included due to the rectangular nature of the identified VOIs. Alternatively, segmentations which more closely follow the boundaries of the heart may be obtained using e.g. graph cuts [115], morphological operations [34] or atlas-based methods [116]. The results showed that for the current application a 3D rectangular bounding box was sufficient. Moreover, this bounding box was successfully acquired in all cases, with an average processing time of 6.8 ± 0.5 s, compared to e.g. 13.2 m reported by [116]. Results presented in [103] illustrate that the method tightly follows a predefined standard anatomical VOI. Finally, the method only required retraining with manually defined 3D bounding boxes in 50 CCTA images.

Similarly to our previous work, x, y, z-coordinates were used to describe the location of each candidate in the image [62]. While in our previous work these features were crucial for accurate scoring, we found that here they only moderately affected the performance of the method. The reason for this may be two-fold. First, the patches provided sufficient texture information for voxel classification. Second, the VOI was sufficiently limited in size that location features did not provide much additional information. Nevertheless, x, y, z-coordinates would likely be valuable to provide artery-specific CAC scores, potentially leading to better prediction of CVD events [78]. It is likely that the proposed method could straightforwardly be ex-

tended to such a multi-class analysis. However, this would increase the complexity of the method and it would therefore require a larger training set size than available in the presented work.

The purely convolutional network architecture used in this study allows training with patches and testing with whole images. To improve efficiency during CNN training, Long et al. proposed end-to-end training with whole images [117], i.e. by minimizing the difference between a predicted and a reference label map for a whole image instead of a single sample. However, the extreme imbalance in our classification problem necessitates balanced sampling of negative and CAC candidates during training. In whole image training, a sampling mask should therefore be applied to the predicted and reference label maps, which would reduce the overlap between patches and the potential increase in efficiency. Furthermore, considering whole images and not patches as samples would reduce the number of possible training batches. Therefore, in this study CNNs were trained with mini-batches containing balanced samples from random training images.

The proposed network architecture (Fig. 5.3) does not include any pooling layers, hence no spatial invariance is introduced. Therefore, the CNNs were trained to predict a label only for the voxel at the center of the odd-sized input patch and not for other voxels in the patch that might have a different label. The experiments showed a benefit of larger input patches in terms of specificity, with input patch sizes of 15 and 25 voxels corresponding to receptive fields of 6.75 mm vs. 11.25 mm along each axis. Note that the typical diameter of a coronary artery is 4 mm [118]. Therefore, a larger patch size provides a wider margin around a coronary artery, which likely allows reduction of FP errors. In this work, only two input sizes were evaluated. As shown by the experiments, smaller inputs would not be likely to provide better results. Larger inputs might provide better results, but this effect might be mitigated by the increase in the number of trainable parameters and the limited number of training samples. Hence, a multiscale approach using a combination of small high resolution input patches to provide detailed local analysis and larger low resolution input patches to provide spatial context might further improve the method, while keeping the number of trainable parameters low.

Our method used either 2.5D or 3D input. We did not use individual 2D planar inputs. Although these are highly efficient, they fail to capture the volumetric aspect of the data. Volumetric 3D patches can provide more information, but their size may pose computational challenges. In our experiments, 3D testing took substantially more time than processing with 2.5D, on average 147 s vs. 52 s for CNN_1 and 201 s vs. 107 s for CNN_2. Similarly, training with 3D input took much longer. In addition, with a limited number of training samples, as is often the case in medical image analysis, 3D CNNs are more likely to overfit to the training data. Our experiments showed a performance drop between 2.5D and 3D architectures in terms of CAC lesion identification in CCTA. It is likely that the number of trainable parameters and input voxels for the 3D patch causes the network to overfit, considering

that the number of parameters was high compared to the number of positive training samples. In other applications in medical image analysis, e.g. [104], 2.5D input outperformed 3D input as in our experiments. As a potential means to overcome the limitations of 3D input patches, yet capture volumetric information, Zheng et al. proposed separable 3D kernels with a reduced number of trainable parameters [119]. Setio et al. extended 2.5D input by using 9 rotated 2D views for CNN-based lung nodule classification, and hypothesized that, to some extent, more 2D views may lead to better performance [120]. However, this advantage may be problem-specific and not applicable to voxel classification. In future work, we will further investigate the trade-off between complexity and performance of different input representations for voxel classification. In addition, the current data set could be enlarged to provide a more diverse set of training samples.

It has previously been shown in 2D natural image classification that ensembles of CNN models can outperform individual models, e.g. in [90]. In our experiments, ensembles of ConvPairs with different dimensionality or input size improved lesion identification in all cases. It is likely that these models captured different aspects of the data, and hence were prone to make different errors. In future work, we will investigate to what extent the combination of models with identical architectures improves results, compared to models with different architectures that were combined in the present study.

A clinical standard for CAC scoring in CCTA is lacking, and intensity thresholds have been determined in various ways [100, 99, 20, 19]. The proposed method was trained using manual annotations in CCTA, based on a patient-specific threshold of $mean_{aorta} + 3SD_{aorta}$ determined with a ROI in the ascending aorta. As a consequence, in several cases, high variability in lumen HU values in the CCTA image caused oversegmentation of the CAC lesion. In the current work, we did not correct for this oversegmentation in the reference. A comparison between automatically determined and manually determined CAC volume and mass scores in CCTA showed that the automatic method generally determined lower scores. In [114], calcium was annotated fully manually by contouring of calcified lesions, an extremely time-consuming process. The performance of our method is likely to increase by using such annotations, i.e. by removing noisy labels from the training data set.

In large studies, CAC scores in CSCT have been shown to be highly predictive of cardiovascular events [8]. For CAC scores in CCTA, such studies have as of yet not been performed. Agatston and volume scores in CCTA are typically much lower in CCTA than in CSCT [114]. CSCT images are reconstructed with 3.0 mm slice spacing, while CCTA images typically have sub-millimeter slice spacing. Therefore, due to partial volume effects high-density CAC appears to have a larger volume in CSCT than in CCTA, but a higher peak intensity in CCTA than in CSCT. The Agatston score uses a stepwise function, which assigns the same weight to all voxels above 400 HU, thus not weighing higher peak intensities in CCTA accordingly. In previous studies, Agatston scores and CAC volume scores in CCTA were converted to modified Agat-

ston scores using empirically determined linear conversion factors [20, 68]. In this study, we primarily evaluated our method using the mass score, which weighs intensities linearly, and therefore better captures differences between CSCT and CCTA. The mass score has previously been shown to yield high correlations between CAC quantification in CSCT and CCTA [121] and to have lower inter-scan variability than the volume and Agatston score [122]. In addition, we did compute unmodified Agatston scores in CCTA to assign patients to CVD risk categories. Although these scores were lower than Agatston scores in CCTA, they ranked patients almost equally, indicating that CAC quantification in CCTA may be used to assign patients to CVD risk categories, without the application of a conversion factor.

CNNs are pattern recognition networks, which means that they learn from examples and hence, are not likely to correctly classify unfamiliar samples. An example of this was found in our test set, which contained large calcified lymph nodes in the pericardium. It is likely that the number of scans in the training set was too small to capture all the variability expected in coronary CT angiography scans. Therefore, in future work a larger set of images covering a larger variety of examples will be included.

The presented method performed very fast voxel classification. Bounding box extraction took on average 7 s per patient. Average processing time for the best performing ConvPair was 46 s for CNN_1 and 28 s for CNN_2. Because our method is fully automatic, it shows potential for application in large-scale studies, as well as in a clinical settings where immediate processing would allow for a smooth workflow. To the best of our knowledge, the relation between CAC scores in CCTA and clinical CVD outcome and/or CVD risk categorization is not yet known. The presented automatic method allowing quick analysis might allow such studies.

In conclusion, CAC can be accurately automatically identified and quantified in CCTA using the proposed pattern recognition method. This might obviate the need to acquire a dedicated CSCT scan for CAC scoring, which is regularly acquired prior to a CCTA, and thus reduce the CT radiation dose received by patients.

Chapter 6

Coronary centerline extraction using simultaneous classification and regression in a single CNN

Based on:
J.M. Wolterink, T. Leiner, R.W. van Hamersvelt, M.A. Viergever, I. Išgum "Coronary Centerline Extraction using Simultaneous Classification and Regression in a Single CNN," *In preparation*

Abstract

Coronary artery centerline extraction from cardiac CT angiography (CCTA) images is a prerequisite for detection of stenoses and non-calcified atheroslerotic plaque. In this work, we propose an iterative tracking algorithm that extracts coronary artery centerlines using a convolutional neural network (CNN).

In the proposed method, a tracker is initialized using placement of a single seed. At each step, a single 3D dilated CNN is used to determine the most likely direction and radius of the artery based on an isotropic patch centered at the tracker's location. Subsequently, the tracker is moved in the identified artery direction, with a step size equal to the identified artery radius. The tracker terminates when no direction can be identified with a predefined certainty level. The CNN is trained using manually defined centerlines in training images. No pre-processing or post-processing is applied, and the process is guided solely by the local image values around the tracker's location.

Eight CCTA images provided in the MICCAI 2008 Coronary Artery Tracking Challenge (CAT08) challenge were used for training and quantitative evaluation, and 50 CCTA images acquired at our own institution (UMCU) were used for validation of the method. In the CAT08 images, the extracted centerlines had an average overlap of 93% with the reference centerlines, and an average distance of 0.27 mm to the centerline points within the reference artery. In the UMCU dataset, 5,448 centerlines were extracted based on seeds placed by an expert. The results showed that most extracted centerlines had substantial overlap with manually placed points. The limits of agreement between reference and automatic radius measurements were below the size of one voxel in both the CAT08 dataset. In the UMCU dataset, centerline extraction required on average 3 s per vessel, with an average vessel length of 164 mm.

The method was able to efficiently and accurately determine the direction and radius of coronary arteries based on information derived directly from the image data.

6.1 Introduction

Accurate information about the geometry and topology of a patient's vasculature is crucial for many medical applications. In patients with suspected coronary artery disease, such information may be obtained non-invasively using a cardiac CT angiography (CCTA) scan [123]. A typical first step in CCTA analysis is the extraction of coronary artery lumen centerlines. This allows multi-planar reconstructions and facilitates stenosis detection and plaque identification. Manual extraction of coronary centerlines is a tedious and time-consuming process, which is infeasible in clinical practice. Therefore, (semi)automatic methods have been proposed for coronary centerline extraction.

In a review on vessel lumen segmentation, Lesage et al. identified different types of centerline extraction methods [124]. First, methods have been proposed to compute a minimum cost path between automatically or manually indicated start- and end-points (e.g. [125]). While methods employing minimum cost paths are robust and typically find high overlap with a reference centerline, their accuracy may be reduced by shortcuts. Hence, it is important to design a cost function that is low at centerlines, and high at other locations. Second, methods have been described that track the centerline based on iterative determination of the vessel's location, orientation and radius [126, 127, 128, 129]. Although such methods are prone to premature stopping at gaps or discontinuities, they have low computational overhead and only explore the CCTA image very sparsely. Third, a number of previously proposed methods first obtain a segmentation [71] or localization [73] of the coronary artery tree, and subsequently recover the line at the center of these structures. Such methods typically do not require any user input, and are likely to be more robust to gaps and discontinuities in the arteries. However, they analyze the full 3D volume and may be time-consuming.

Existing methods for centerline extraction typically use filters or models based on certain assumptions about vessels to determine the coronary location, orientation and radius. Commonly used filters are based on eigenvectors of the Hessian matrix [70] or idealized tubular models of vessels [126]. Such filters and models are hand-crafted and may not fully grasp the information available in the data. Consequently, they require adaptation to cases in which the underlying assumptions do not hold. Such adaptations may be required at coronary branching points, or to suppress responses in non-coronary structures [71]. Explicit modeling of all exceptions is a challenging task.

More recently, machine learning methods have been proposed that learn vessel models from annotated data. Gülsün et al. proposed a method in which a vessel orientation flow field is determined by a support vector machine (SVM) classifier using geometric features based on a heart-oriented coordinate system and steerable features [130]. Schneider et al. [131] proposed to use steerable features and randomized decision trees to locate centerlines by voxel voting, while Sironi et al. [132] trained a boosting regressor to predict, for each voxel, the displacement to the clos-

est centerline. In both [131] and [132], the final centerline was extracted by finding a minimum energy path in the resulting image using a fast marching algorithm. Although these methods do not depend on hand-crafted vessel filters or models, they require evaluation at multiple voxel locations to obtain robust centerline results.

Convolutional neural networks (CNNs) have previously demonstrated the ability to derive useful features from image data in a wide range of medical image analysis tasks, among which tasks involving vessels (e.g. [133, 134, 135]). This suggests that CNNs could also be used for coronary centerline extraction. In this work, we propose a CNN that learns to identify the coronary centerline direction and lumen radius directly from image data only. The method is trained with manually annotated reference centerlines, and simultaneously predicts the most likely centerline direction and vessel radius at any location in or near the coronary artery based on a local 3D isotropic image patch. Hence, all information is extracted directly from the image, and no intermediate vesselness representations are required. We evaluate the performance of the proposed method using the publicly available Rotterdam Coronary Artery Evaluation Framework[1] and a set of CCTA scans acquired at our own institution.

6.2 Data

6.2.1 CAT08 dataset

Schaap et al. presented an independent comparison of coronary centerline extraction methods in CCTA in the MICCAI 2008 Coronary Artery Tracking Challenge (CAT08), part of the Rotterdam Coronary Artery Evaluation Framework [54]. CAT08 provides 32 CCTA scans, acquired on a 64-slice CT scanner (Sensation 64, Siemens Medical Solutions, Forchheim, Germany) or dual-source CT scanner (Somatom Definition, Siemens Medical Solutions, Forchheim, Germany). Scans were acquired with 120 kVp and a maximum tube current of 900 mA. Images were reconstructed to a mean voxel size of $0.32 \times 0.32 \times 0.4$ mm^3. In each scan, the centerline and radius of four major arteries were manually annotated in a consensus reading by three experts. These annotations are provided for 8 training scans, while the other 24 scans are considered test scans. A detailed description of scan acquisition and reconstruction, and the centerline annotation protocol is provided in [54].

6.2.2 UMCU dataset

We included 50 CCTA scans that were consecutively acquired at our institution (University Medical Center Utrecht, Utrecht, The Netherlands). The need for informed consent was waived by the local medical ethics committee. These scans were acquired with a 256-detector row Philips Brilliance iCT scanner, with contrast enhancement and ECG-triggering, using 120 kVp and 210-300 mAs. Images

[1] http://coronary.bigr.nl/centerlines

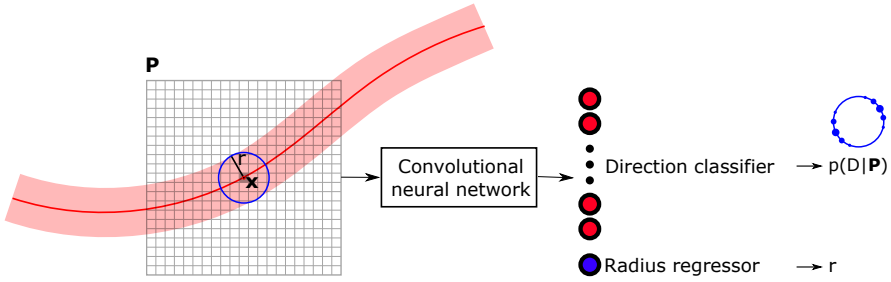


Figure 6.1: Overview of the proposed method. At location **x**, a 3D patch **P** is extracted and used as input to a convolutional neural network (CNN). This CNN simultaneously determines a probability distribution $p(D|\mathbf{P})$ over a discrete set of directions on a sphere (here shown as a blue circle), and an estimate r of the radius of the vessel.

were reconstructed to a mean voxel size of $0.45 \times 0.45 \times 0.45$ mm^3. The quality of the images was generally good. Intravascular stents were present in 2/50 scans, step reconstruction artifacts in 7/50 scans, and coronary motion artifacts in 11/50 scans. Furthermore, 26/50 scans contained coronary artery calcification. For those patients with coronary artery calcification, Agatston calcium scores in corresponding non-contrast CT images ranged from 0.5 to 3,540, with a median (IQR) of 31.6 (6.3–610.5) [9]

In each of the 50 scans, an expert observer manually annotated coronary centerline points separated by approximately 1 cm, along with a measurement of the vessel radius in the axial plane. Seeds were placed in the major coronary arteries, as well as all visible branches. The radius of the placed seedpoints ranged from 0.45 mm to 3.51 mm, with a median (IQR) of 0.81 (0.61–1.16) mm. In total, 5,448 reference points were placed, for an average of 109 points per CCTA scan.

6.3 Method

We propose a single CNN that simultaneously determines the orientation and radius of a coronary artery at a location **x** in 3D CCTA volume (Fig. 6.1). The output layer of the CNN consists of classification nodes that determine a posterior probability distribution over possible tracking directions D, and a regression node that determines the radius r of the vessel at point **x**. These values are based only on an isotropic 3D image patch **P** centered at **x**.

6.3.1 Convolutional neural network

The proposed CNN takes a 3D image patch **P** with width w and isotropic voxel spacing v centered at **x** as input, and determines a posterior probability distribution $p(D|\mathbf{P})$ over a discrete set of possible directions D, as well as an estimate r of the

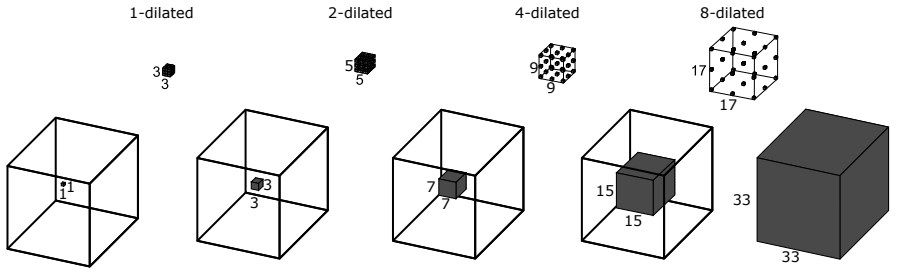


Figure 6.2: Stacked application of 3D convolution kernels with increasing levels of dilation. A $3 \times 3 \times 3$ voxel kernel is shown with a stride of 1, 2, 4 or 8 voxels between kernel elements. These kernels rapidly increase the receptive field from $1 \times 1 \times 1$ voxel to $33 \times 33 \times 33$ voxels using only 27 trainable parameters per kernel.

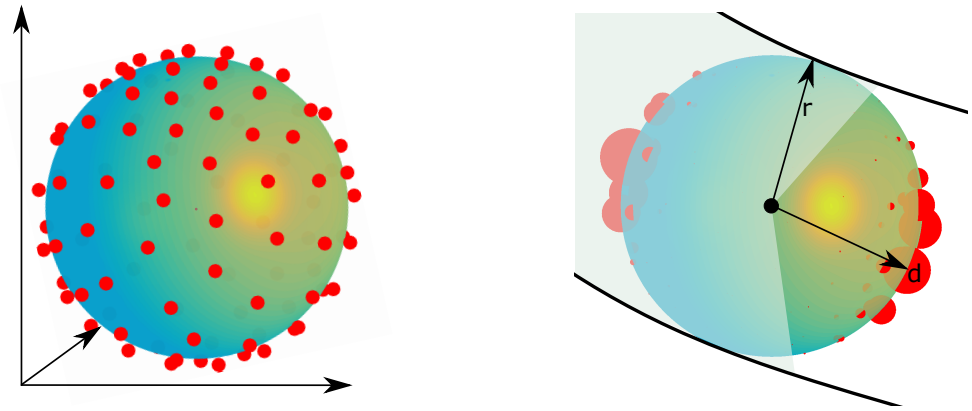
Table 6.1: Architecture of the convolutional neural network (CNN) for a $35 \times 35 \times 35$ voxel 3D input patch. For each layer (Layer), the convolution kernel width (Kernel width) is listed, as well as the dilation level (Dilation), the number of output channels (Channels), and the receptive field at that layer (Field width). All operations are performed in 3D. The CNN can be adapted to $19 \times 19 \times 19$ voxel input patches by omission of layer 5.

Layer	1	2	3	4	5	6	7	8
Kernel width	3	3	3	3	3	3	1	1
Dilation	1	1	2	4	8	1	1	1
Channels	32	32	32	32	64	128	256	$\lvert D\rvert+1$
Field width	3	5	9	17	33	35	35	35

radius. The values for w and v together determine the input resolution and physical receptive field of the CNN in world coordinates.

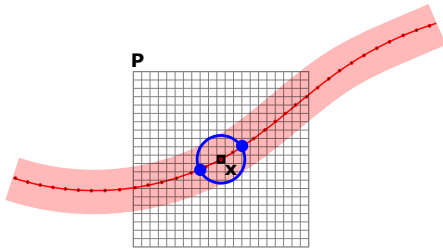
Direction and radius predictions at each location should be accurate and localized. Therefore, the proposed CNN architecture does not use pooling layers, which typically introduce translation invariance. Instead, to compress 3D patch $\mathbf{P}$ into a feature vector representing its characteristics, a stack of dilated convolution layers is used. This stack aggregates features over multiple scales using convolution kernels with increasing levels of dilation, i.e. increased strides between the kernel elements [136]. Fig. 6.2 shows the effect of stacking $3 \times 3 \times 3$ voxel 3D convolution kernels with increasing levels of dilation (1, 2, 4, and 8). As the level of dilation increases, the receptive field of each kernel increases, but the number of trainable parameters per kernel stays the same at $3 \times 3 \times 3 = 27$. While the receptive field grows exponentially, the number of trainable parameters increases linearly. Reducing the number of parameters could prevent overfitting in 3D CNNs.

Table 6.1 lists characteristics of the proposed CNN architecture; the width of convolution kernels, the level of dilation, the number of output channels, and the receptive field at each layer. The CNN can process 3D patches of $w = 35$ voxels, or

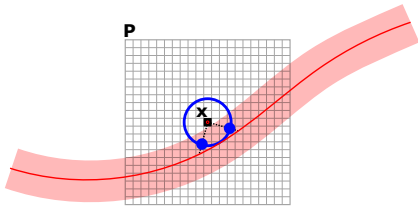


(a) (b)

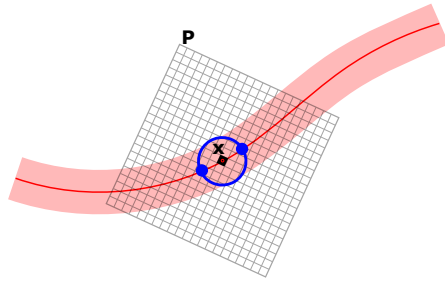
Figure 6.3: (a) The directions D are distributed on a sphere. (b) During testing, a posterior probability distribution over D is determined, and the tracker follows the direction d corresponding to the maximum in this distribution. Only directions with an angle $\leq 60°$ to the previously followed direction are considered.



(a) Training sample



(b) Translated training sample



(c) Rotated training sample

Figure 6.4: Training samples for the convolutional neural network (CNN). (a) Standard training sample extracted at point x, consisting of a patch $\mathbf{P}$ and two reference directions in D (blue circle). Manually defined reference points are shown on the centerline. During training, a random combination of translation and rotation augmentations is applied to each sample. (b) A translated off-centerline point x with corresponding patch $\mathbf{P}$ and reference directions in D. (c) A centerline point x with a rotated patch $\mathbf{P}$ and reference directions in D.

patches of $w = 19$ voxels when layer 5 is omitted. No dilation is applied in the first two or final three layers of the network [136]. To convey more information about an increasing receptive field, the network gets wider towards the output layers.

The output layer combines two tasks: direction classification and radius regression. The possible directions D are distributed on a sphere, where each point corresponds to a class (Fig. 6.3a). We use classification instead of regression, so that the CNN may return a posterior probability distribution with multiple local maxima (Fig. 6.3b). In contrast, a regression model minimizing the squared error between the predicted and reference direction would predict the average direction, which in most cases is the center of the sphere. The $|D|$ classification nodes are combined through a softmax activation layer. Radius regression is performed in the CNN's output layer using a single output node with a linear activation function.

Aside from the classification and regression nodes in the output layer, all nodes in the network use rectified linear units as an activation function. Batch normalization is applied in each layer of the network [137]. In addition, to prevent overfitting, dropout is applied before layers 7 and 8, with a probability 0.5 [113]. Fully connected layers are implemented as $1 \times 1 \times 1$ convolutions. Hence, after training the network may efficiently be applied to images of arbitrary size.

6.3.2 Training strategy

The CNN is trained with CCTA scans for which reference centerlines are available. This reference standard consists of an ordered set of centerline points with corresponding radius measurements. To generate a training sample, a point $\mathbf{x}$ on the reference centerline is randomly selected and a 3D image patch $\mathbf{P}$ centered at that point is extracted from the training image (Fig. 6.4a). The reference directions are determined as follows. First, the displacement between $\mathbf{x}$ and a point at a distance r along the centerline is determined, where r is the reference radius. The directions in D are normalized to length r and the direction $d \in D$ (Fig. 6.3a) that best matches the displacement is identified. This direction is considered one of the two reference directions at point $\mathbf{x}$ and its class label is set to 0.5 in the reference distribution. Because there are two reference directions, this process is repeated with the reversed centerline. Hence, two classes in the reference distribution have value 0.5 and all other probabilities are set to zero. The patch $\mathbf{P}$, the reference distribution, and the reference radius r together form an input sample for the CNN.

Training with a large and diverse set of training samples is likely to improve the performance of the proposed supervised machine learning method. However, obtaining reliable reference centerlines in a large number of images is tedious and time-consuming, and may require consensus of multiple experts [54]. Hence, we augment the already available reference centerline data in two ways.

First, we include samples that are located off the coronary artery centerline. Training with only samples on the vessel centerline could lead to drifts from the centerline, from which the tracker may be unable to recover. The location of an off-

centerline sample $\mathbf{x}$ is chosen from a 3D normal distribution centered at a random reference centerline point with $\sigma = 0.25r$ (Fig. 6.4b). To determine the reference displacement from an off-centerline point, we first identify the closest reference point. We then find the point at a distance r along the centerline and let the reference direction from $\mathbf{x}$ be in the direction of this point, normalized to reference radius length r. This is repeated for the reverse vessel direction.

Second, we enrich the dataset by applying random rotations to input patches (Fig. 6.4c). Each training patch $\mathbf{P}$ is rotated around the x-, y-, or z-axis with a random angle $\alpha \in [0, 2\pi)$ and the inverse rotation is applied to the training distribution. This balances the orientation of vessels in the training set and makes the CNN agnostic to the orientation of the image.

During training, the Adam optimizer [138] updates the network parameters θ to minimize the loss

$$\ell(\theta) = \ell_c(\theta) + \lambda_r \ell_r(\theta) + \lambda_w \|\theta\|^2, \tag{6.1}$$

where ℓ_c is the categorical cross-entropy between the reference and posterior probability distribution over the direction classes, $\lambda_r \ell_r$ is the squared error regression loss between the reference and predicted radius values, weighted by a parameter λ_r, and $\lambda_w \|\theta\|^2$ is a regularization term on the network parameters. We used $\lambda_r = 15$ and $\lambda_w = 0.001$ throughout our experiments. Mini-batch training is used with batches containing 64 randomly selected samples.

6.3.3 Iterative tracking

The CNN architecture used is fully convolutional and able to process input images of any size. Hence, a trained CNN may either be used to compute the vessel orientation and radius in all voxels in a full-sized image prior to centerline extraction, or to guide a tracker that iteratively processes locally extracted 3D patches. While the first method is likely to be more robust to occlusions or artifacts in the vasculature, the second approach has a smaller computational footprint. In addition, precomputing is done with an isotropic discrete grid, while iterative tracking uses world coordinates and is thus potentially more accurate.

We here demonstrate the ability of the CNN to determine accurate centerline directions using an iterative tracking algorithm. The tracker starts at a seed point $\mathbf{x}_0$. An isotropic 3D patch $\mathbf{P}_0$ is extracted at $\mathbf{x}_0$ using trilinear interpolation and processed by the CNN. Two initial directions d_0 and d_0' separated by an angle $\geq 90°$ are identified as local maxima in the posterior probability distribution $p(D|\mathbf{P}_0)$. The tracker first extracts the centerline in the direction d_0. The tracker moves to point $\mathbf{x}_1 = \mathbf{x}_0 + r_0 \frac{d_0}{\|d_0\|}$, where r_0 is the estimated radius at point $\mathbf{x}_0$. This process is repeated until a stopping criterion is fulfilled. Subsequently, the tracker follows the same process in the direction d_0', starting again at point $\mathbf{x}_0$.

The stopping criterion is based on the uncertainty of the direction classifier. At every point, the normalized entropy $H(p(D|\mathbf{P})) \in [0, 1]$ in the posterior probability

distribution is computed as

$$H(p(D|\mathbf{P})) = \frac{\sum_{d\in D} -p(d|\mathbf{P}) \log_2 p(d|\mathbf{P})}{\log_2 |D|}. \tag{6.2}$$

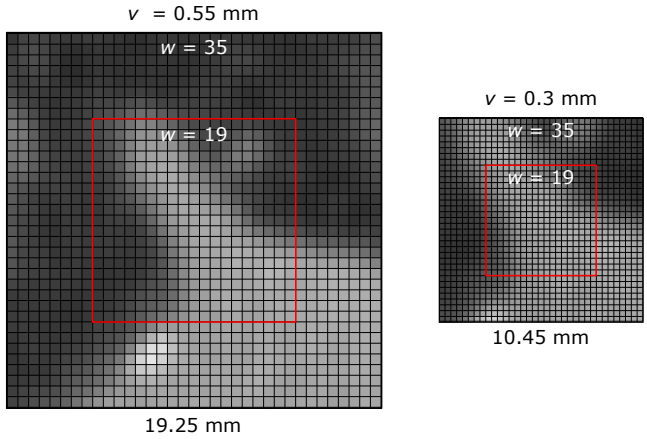
If this entropy crosses a threshold of 0.5, i.e. if the probability distribution over the possible directions is too homogeneous, the tracker first tries to obtain more information from randomly rotated 3D patches centered at the current location, similar to random view aggregation proposed by [105], and selects the posterior probability distribution of the rotation that results in the lowest entropy. If the entropy of the selected probability distribution crosses a threshold of 0.7, the tracker terminates. This may be the case when the end of the artery is reached, or when the tracker encounters a non-coronary-artery structure. In addition, stenotic areas or areas with low image contrast may ocassionally lead to high entropy values. To encourage tracking through such areas, the termination entropy is determined as a moving average over the past few steps, similar to the probabilistic tracking scheme proposed by Wang et al. [139]. Finally, to prevent the tracker from following a path that overlaps with its already extracted centerline, tracking is terminated when the tracker is at a point which is too close to a previously tracked point.

6.4 Experiments and results

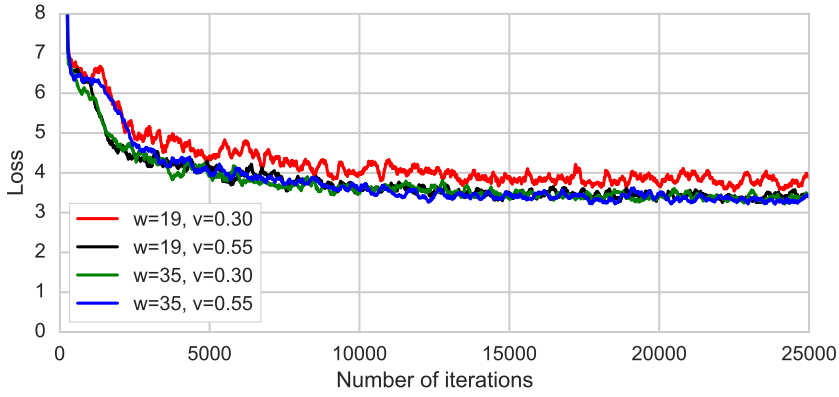
6.4.1 Parameter selection

To determine the optimal size and resolution of input patches to the CNN, we performed experiments with different values for patch width w and voxel size v, namely $w = \{19, 35\}$ voxels and $v = \{0.3, 0.55\}$ mm (Fig. 6.5a). The values for w and v were chosen so that a $(w = 35, v = 0.3)$ patch and a $(w = 19, v = 0.55)$ have an equivalent physical receptive field, spanning the typical coronary artery diameter with added margin.

We trained a CNN using seven training images and one validation image from the CAT08 dataset. Fig. 6.5b shows the validation loss for the four different input configurations. From these results it can be observed that using smaller patches with a lower resolution ($w = 19$, $v = 0.55$ mm) did not lead to an increase in validation loss compared to large patches with high resolution ($w = 35$, $v = 0.30$ mm). However, using small patches with a high resolution ($w = 19$, $v = 0.30$ mm) did lead to reduced performance. Finally, larger patches with lower resolution $(w = 35, v = 0.55)$ only provided limited improvement. In addition, we found that a CNN with $w = 35$ required around eight times more time than a CNN with $w = 19$. Hence, all subsequent experiments were performed with $w = 19$, $v = 0.55$. In all experiments, the number of directions in the set D was set to 200. In case of uncertainty of the direction classifier ($H(D|\mathbf{P}) > 0.5$), up to ten additional randomly rotated patches were evaluated.



(a) Example patch widths and voxel sizes



(b) Validation loss for different patches

Figure 6.5: (a) Illustration of different patch sizes w and voxel sizes v. In all cases, the smaller red rectangle shows $w = 19$ and the larger rectangle shows $w = 35$. A lower value for v leads to a higher resolution input image, but with a reduced physical receptive field. A high value for v substantially enlarges the physical receptive field, but leads to a loss in granularity. (b) Experiments using different patch sizes and voxel spacings show that the smallest high resolution patch leads to the highest validation loss.

The algorithm was implemented in Theano [111] and Lasagne [140]. In our implementation, the algorithm performed most computing on the GPU. We analyzed the required computing time per vessel using a NVIDIA Maxwell Titan X GPU with 3072 CUDA cores and 12 GB memory, as well as using a standard workstation's NVIDIA Quadro K2000 GPU with 284 CUDA cores and 2 GB memory. We found that the time required for extraction of a single centerline using the NVIDIA Maxwell Titan X GPU was on average 3 ± 1.2 s. This corresponds to a normalized tracking speed of 5 cm/s. In comparison, we found that tracking using the NVIDIA Quadro K2000 GPU tracking was approximately two times slower, but still feasible.

6.4.2 Centerline extraction

CAT08 dataset

We evaluated centerline extraction performance on the eight training scans provided in CAT08 using a leave-one-out cross-validation; for each training scan, a CNN was trained with 100,000 mini-batches using samples from the other seven training scans. To extract a centerline, we manually placed a single seed point that uniquely identified the coronary artery.

Table 6.2 lists results for each of the 32 coronary arteries in the training set, showing total overlap (OV), overlap until first error (OF), overlap of the extracted centerline with the clinically relevant part of the vessel (OT), and average inside accuracy (AI) of the automatically extracted vessels. The last metric measures the average distance between the reference and extracted centerline for points that are within the radius of the reference centerline. We found that the iterative tracker tended to give the strongest response in the direction of larger vessels leading to the ostium, and reached the coronary ostium in all cases. The average total overlap with the reference centerline was 93%, with a minimum of 73% and a maximum of 100%. The average overlap until the first error was 82%, with a minimum of 12% and a maximum of 100%. The average overlap with the clinically relevant centerline (radius ≥ 0.75 mm) was 95%, with a minimum of 73% and a maximum of 100%. The average inside accuracy was 0.27 mm, which is smaller than typical voxel width in the dataset.

Reduced centerline overlap was caused by either undersegmentation or oversegmentation. In one case, coronary artery bridging caused premature termination of the tracker. In a second case, the coronary artery was narrow, had very low contrast with the surrounding tissue and crossed a vein with higher attenuation than the artery. Because the tracker only follows a single hypothesis, it partially followed this vein before terminating. In several cases, the tracker partially followed a connected or adjacent vein or artery segment after reaching the artery's most distal reference point.

Table 6.2: Centerline extraction overlap and accuracy results in the CAT08 training set, obtained with leave-one-image-out cross-validation. For each vessel, the total overlap (OV), overlap until first error (OF), clinically relevant overlap (OT) and inner accuracy (AI, in mm) are listed. In addition, the average over all vessels is shown.

Image	Vessel	OV	OF	OT	AI
0	0	0.80	0.12	0.80	0.49
	1	0.86	0.84	0.86	0.34
	2	0.94	0.40	0.94	0.31
	3	0.93	1.00	1.00	0.22
1	0	0.99	1.00	1.00	0.32
	1	0.98	1.00	1.00	0.24
	2	0.97	0.95	1.00	0.30
	3	0.84	0.96	0.89	0.28
2	0	0.94	1.00	1.00	0.24
	1	1.00	1.00	1.00	0.24
	2	0.97	1.00	1.00	0.27
	3	0.99	1.00	1.00	0.31
3	0	0.93	1.00	1.00	0.25
	1	0.73	0.20	0.74	0.26
	2	0.90	0.94	0.95	0.17
	3	0.93	0.36	0.98	0.27
4	0	0.99	0.98	1.00	0.19
	1	0.91	0.23	0.91	0.21
	2	0.83	0.84	0.88	0.24
	3	0.98	1.00	1.00	0.18
5	0	0.92	1.00	1.00	0.37
	1	0.93	0.87	0.93	0.26
	2	1.00	1.00	1.00	0.31
	3	0.97	1.00	1.00	0.25
6	0	0.96	0.99	1.00	0.28
	1	0.99	1.00	1.00	0.20
	2	0.92	0.94	0.99	0.29
	3	1.00	0.99	1.00	0.18
7	0	0.95	0.90	1.00	0.26
	1	0.86	0.28	0.86	0.28
	2	0.73	0.58	0.73	0.30
	3	1.00	1.00	1.00	0.24
Average		0.93	0.82	0.95	0.27

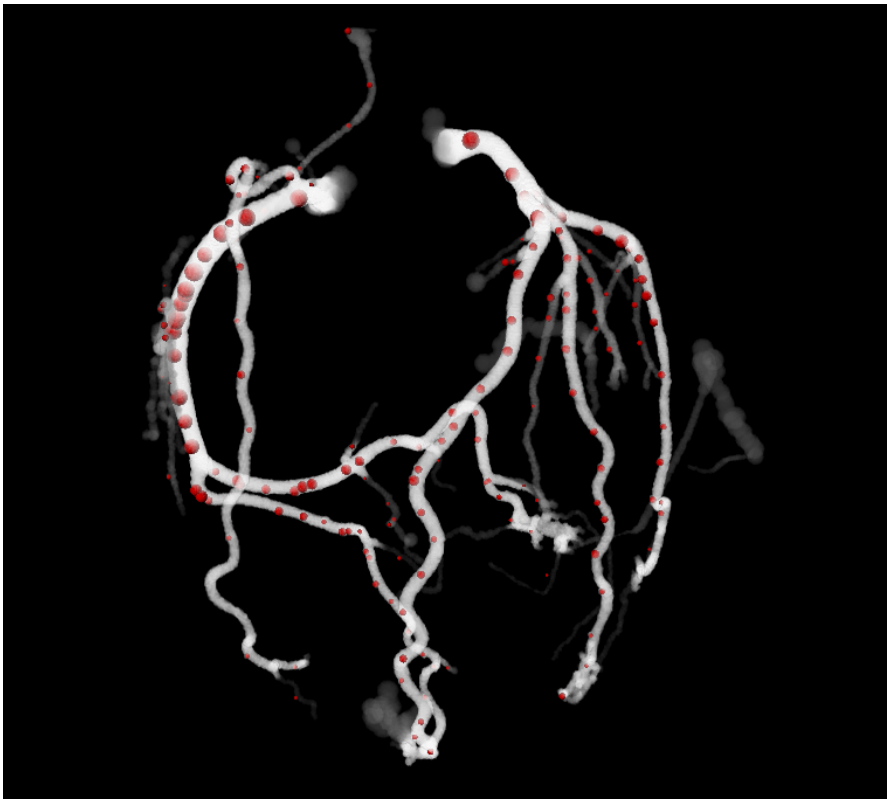


Figure 6.6: Volume rendering of extracted centerlines with estimated radius. Red spheres indicate manual reference seed points set by a human observer. The opacity at each voxel indicates the number of centerlines in the vicinity.

UMCU dataset

To assess whether training with scans acquired on a Siemens CT scanner in the CAT08 dataset allows testing with the UMCU set acquired on a Philips CT scanner in our own hospital, we retrained a single CNN using the eight CAT08 training scans and 250,000 mini-batch iterations. Although the detail of annotation in the UMCU dataset is lower than in the CAT08 dataset, the number of images is larger. We initialized tracking in each of the 5,448 manually placed seed points. Fig. 6.6 shows an example of seed points placed by the human expert and all centerlines extracted from those seed points, dilated by their estimated radii. The vessel opacity at each point corresponds to the number of centerlines near that point. Hence, as can be expected, most centerlines pass through the proximal coronary arteries, and fewer centerlines pass through coronary artery branches.

Fig. 6.7 shows a histogram of the lengths of 5,448 extracted centerlines in the UMCU dataset. The histogram shows that the vast majority of centerlines had a

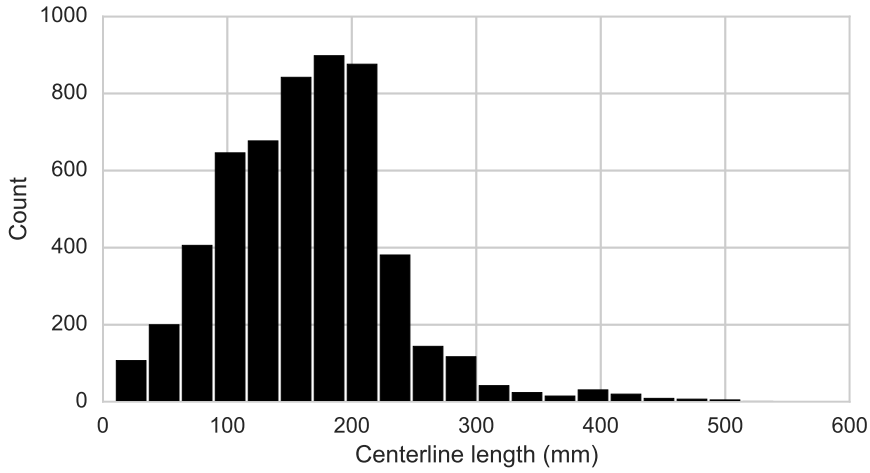


Figure 6.7: Length of 5,448 extracted centerlines based on reference points in the UMCU dataset containing 50 scans.

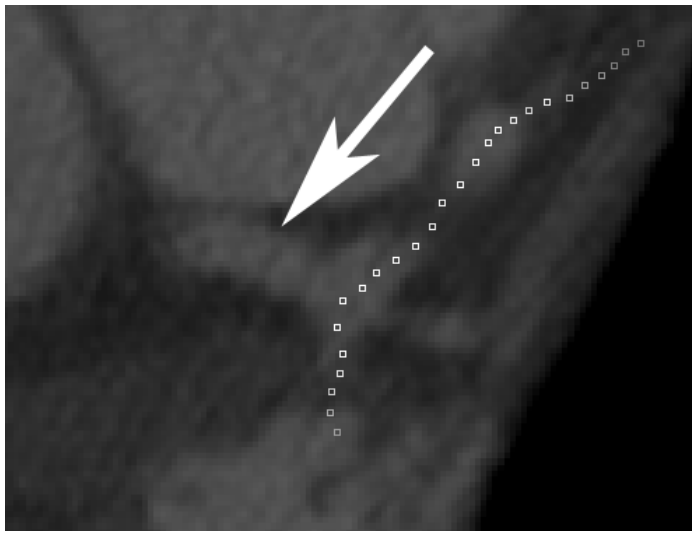


Figure 6.8: An example where the tracker (white squares) continued into a branch instead of following the path to the ostium (white arrow).

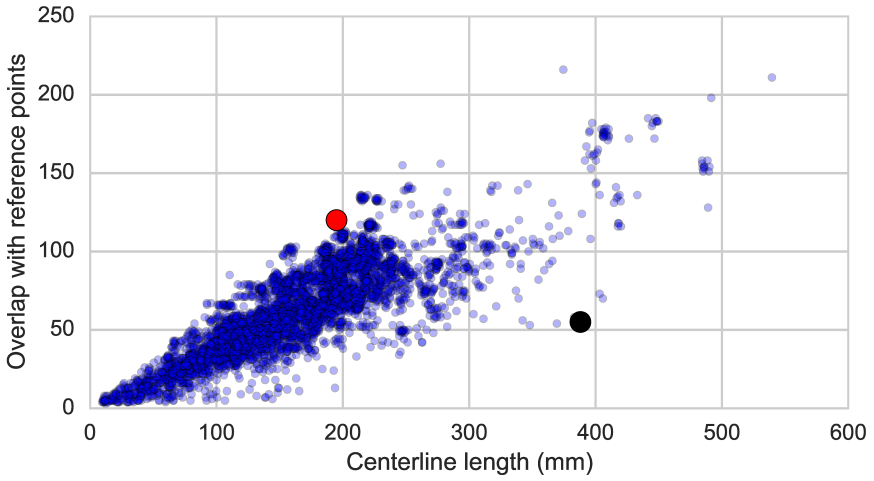


Figure 6.9: Number of overlapping points with the manual reference as a function of the length of the artery in 5,448 centerlines. The red marker indicates a properly tracked centerline, the black marker a centerline that erroneously entered a vein.

length of less than 200 mm, which is typical for coronary arteries [141]. In long extracted centerlines closer inspection revealed that the tracker's termination criteria were sometimes not sufficient, causing the tracker to continue into another nearby coronary artery after reaching the end of the artery. In a number of cases, the tracker made a loop from the left coronary ostium to the right coronary ostium by extracting the centerlines of both arteries and connecting these through a short segment of a vein. In several other cases, extracted centerlines were longer because the tracker did not find the ostium, but instead followed a different artery at a branching point. Fig. 6.8 shows an example of such a case, where the tracker was seeded in the left anterior descending coronary artery and continued into the left circumflex branch instead of the left main branch leading to the coronary ostium (indicated by a white arrow). On the other hand, the histogram shows some short centerlines. These were typically due to seeding in short coronary branches, or because an artifact was encountered.

To further investigate the performance of the tracker, we computed the overlap between each extracted centerline and the manual reference points. This overlap was determined as the number of extracted centerline points that pass through the reference radius of a manually placed point. Fig. 6.9 shows a scatterplot of this metric as a function of vessel length, in which each extracted centerline is a data point. In general, centerlines with high overlap indicate good vessel tracking. An example of this, indicated in red in Fig. 6.9, correctly followed the coronary artery from the ostium down to a radius of 0.66 mm. Points below the cloud indicate poorer vessel tracking: in one case, indicated in black in Fig. 6.9, the scan contained

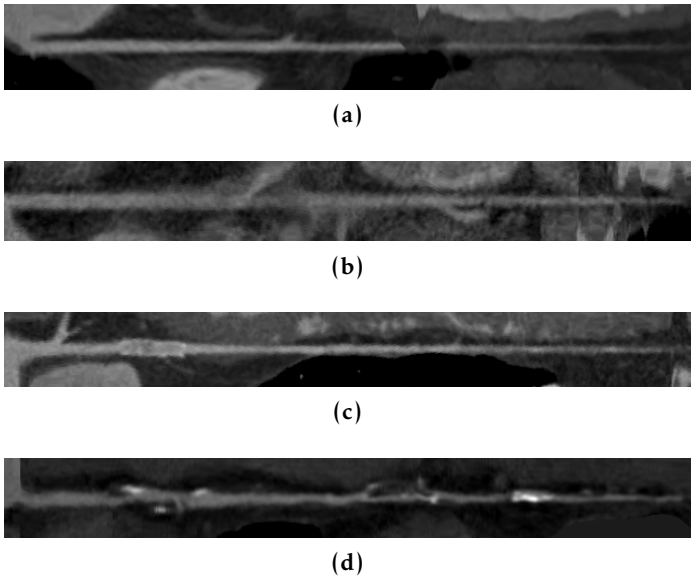


Figure 6.10: Stretched multi-planar reconstructions of successfully extracted coronary artery centerlines in the presence of a (a) step artifact, (b) motion artifact, (c) intravascular stent, and (d) coronary calcification.

a large step artifact and veins with relatively high attenuation. At the step artifact, the tracker left the artery and continued to track the coronary vein, resulting in a large false positive segment.

We found that the performance of coronary artery centerline extraction can depend on the imaging protocol, the quality of the reconstructed image and characteristics of the patient. In general the image quality in the UMCU dataset was higher than in the CAT08 set, due to a higher contrast between the lumen and the surrounding tissue. In some cases, attenuation in the veins was also high, making it difficult for the CNN to distinguish between coronary arteries and coronary veins. Nevertheless, the method was able to extract the centerline in cases with small step artifacts (Fig. 6.10a), mild motion artifacts (Fig. 6.10b), in patients with intravascular stents (Fig. 6.10c) and in patients with coronary artery calcification (Fig. 6.10d). Because the CNN is to some extent invariant to local rotations of the vessel, and uses a relatively small physical receptive field, it makes very few assumptions about the anatomy around an artery. This makes the method useful in cases with abnormal anatomy. For example, Fig. 6.11 shows a patient with three tortuous left coronary artery branches that are correctly extracted based on manually placed seed points.

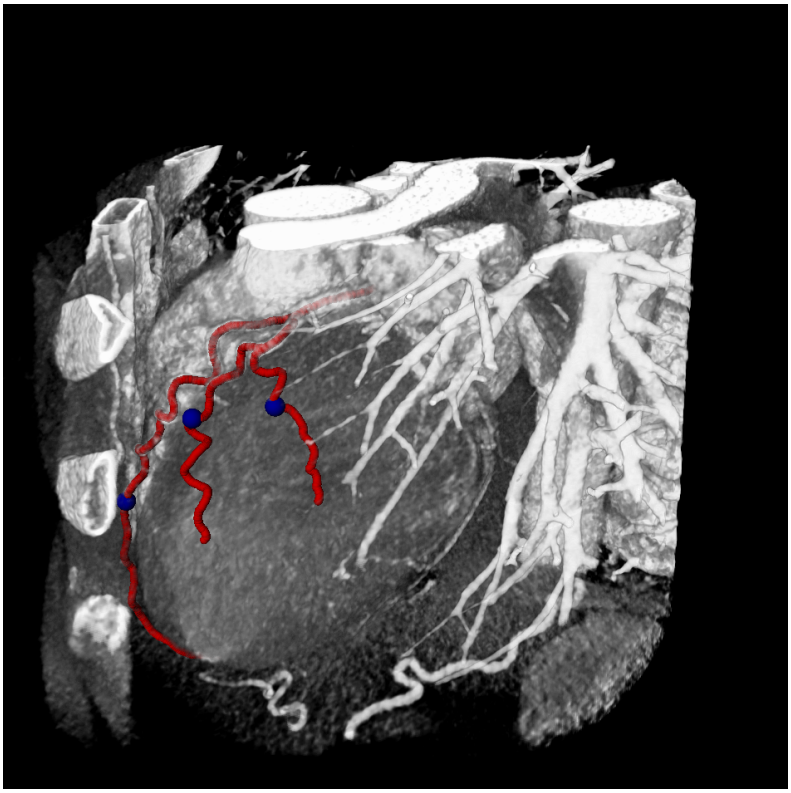


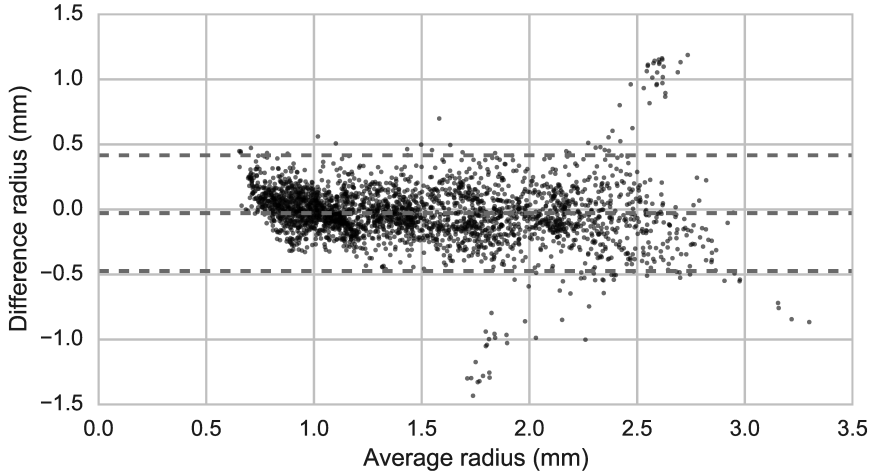
Figure 6.11: Successful centerline extraction in a patient with tortuous coronary arteries. Three seed points provided by an expert are shown in blue and extracted centerlines are shown in red.

6.4.3 Radius estimation

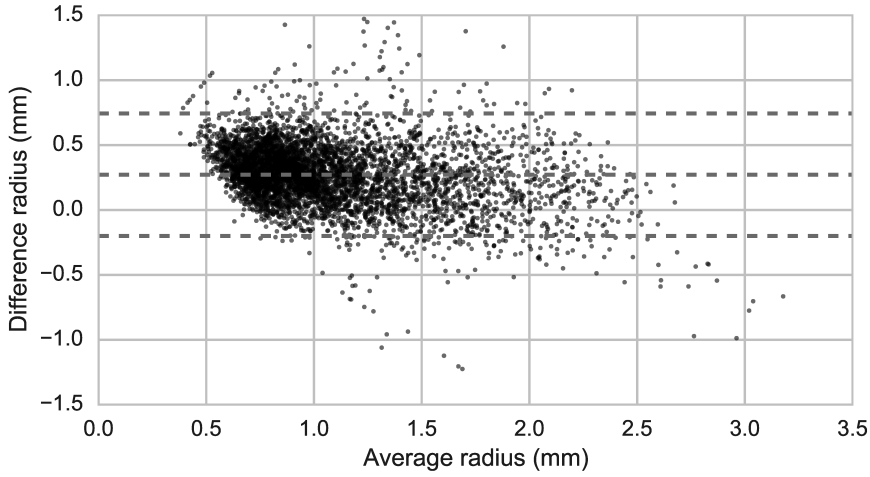
The regressor output node of the CNN estimates the vessel radius at a location $\mathbf{x}$ in the coronary artery. We assessed how well these radius estimates correspond to reference radius values.

CAT08 dataset

We used the models trained during cross-validation to compute the estimated radius value at every reference centerline point in the CAT08 training set and compared these to the reference radius values. The Bland-Altman plot in Fig. 6.12a shows that our method had a very slight systematic underestimation of -0.03 mm. Nevertheless, the limits of agreement (−0.47–0.42) mm were within the width of a typical CCTA image voxel.



(a) CAT08 dataset



(b) UMCU dataset

Figure 6.12: Bland-Altman plot comparing reference and automatically determined radius values. The y-axis shows the difference between the automatically determined and manually determined radius values. (a) 139,337 centerline points along 32 coronary arteries in the CAT08 dataset. For visualization purposes, every 50th point is shown. Values were obtained using leave-one-image-out cross-validation. (b) 5,448 centerline points in the UMCU dataset.

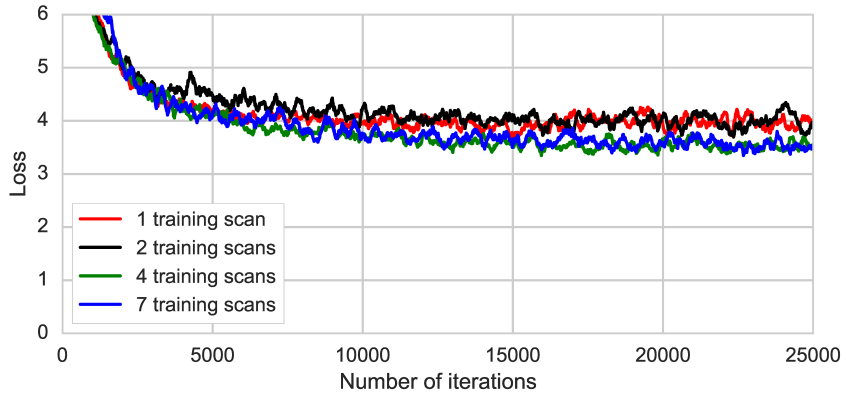


Figure 6.13: Validation loss on experiments using different numbers of training scans.

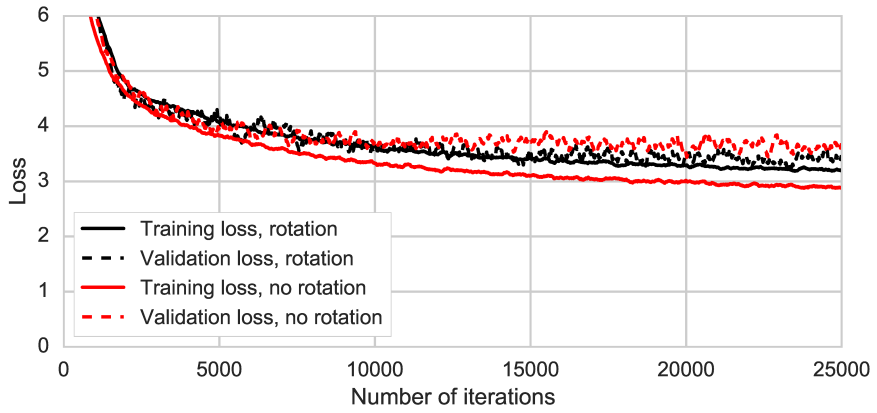


Figure 6.14: Effect of rotation data augmentation on learning curves for the training and validation set. Without rotation augmentation, the training loss rapidly decreases, while the validation loss increases after initially decreasing.

UMCU dataset

In addition to the CAT08 scans, we evaluated radius prediction performance at every point in the UMC scans indicated by the observer. Fig. 6.12b shows a Bland-Altman plot comparing automatic and reference radius values. While the spread of the differences is similar to the CAT08 dataset, there was a systematic overestimation (0.27 mm) by the automatic method compared to the manually annotated reference. A potential reason for this may be caused by differences in annotation protocol. The manual radius values in the UMCU dataset were determined in the axial plane, while in the CAT08 dataset they were determined in cross-sectional reconstructions [54]. Therefore the CNN may have learned to estimate radius values differently than the expert. A difference in radius values was also observed when comparing the distribution of radius values in the CAT08 dataset (median [IQR] 1.37 [1.03–1.88] mm) with the distribution of radius values in the UMCU dataset (median [IQR] 0.81 [0.61–1.16] mm).

6.4.4 Method analysis

We performed several experiments to further investigate the characteristics of the method.

Amount of training data

Acquisition of a reference centerline for a coronary CT angiography scan is a time-consuming process. Therefore, the number of training scans with reliably delineated centerlines is limited. To investigate how much training data the method requires, we trained CNN networks with 1, 2, 4 or 7 training scans, each time using the same scan for validation. Fig. 6.13 shows the validation loss for each of these networks. From this we can observe that the loss drops substantially when using 4 instead of 1 or 2 training scans. However, extending the training set to 7 scans does not seem to have a large influence on the validation loss.

Data augmentation

To evaluate the effect of data augmentation during training, we trained an additional model without such augmentation. Fig. 6.14 shows a comparison between this CNN and the CNN trained with rotated samples. The learning curves show that the network without rotated samples overfits rapidly, i.e. the validation loss plateaus while the training loss continues to drop. In contrast, when rotated samples are included, the training and validation loss are much more similar and decrease simultaneously. Note that in neither case augmentation is applied to the validaton samples.

Likewise, to evaluate whether the inclusion of off-centerline training samples affects the performance of the CNN, we train one model with and one model without off-centerline augmentation. We selected one image and initialized tracking from

a seed point. This point was randomly translated 50 times. Fig. 6.15 shows that the CNN trained without translation augmentation is extremely sensitive to seed placement, and unable to recover from deviations from the centerline. In contrast, the CNN trained with translation augmentation immediately recovers the centerline from erroneous seed placement. In addition, the centerlines extracted from randomly shifted seed points are very similar, and most reach the starting and endpoint of the coronary artery.

Combining classification and regression

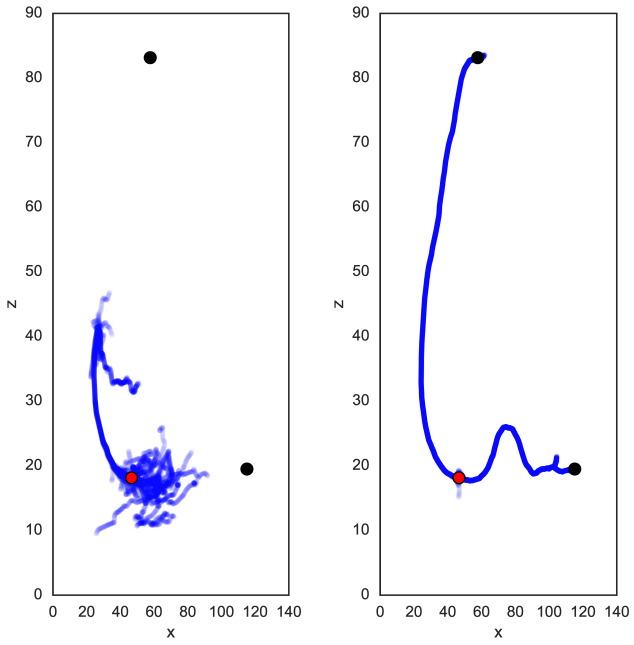
The proposed CNN is trained to simultaneously perform direction classification and radius regression. To investigate to what extent the combination of these tasks affects the performance of the CNN on either of the tasks, we trained three CNNs. The first CNN was trained to only perform classification, the second CNN was trained to only perform regression, and the third CNN was trained to perform both tasks simultaneously. Fig. 6.16 shows the evolution of the classification loss ℓ_c and the regression loss ℓ_r during training of these three CNNs. The results show that although the CNNs trained to perform only classification or only regression converge faster, the CNN combining both tasks ultimately converges to the same loss values.

6.5 Discussion

We have presented a deep learning-based method for coronary artery centerline extraction in CCTA. The method does not require hand-crafted filters or features to determine the orientation and size of the coronary artery, but instead uses a CNN to directly extract this information from the available CCTA image data.

Our experiments showed that the CNN was able to provide a single-vessel seed-based iterative tracker with information about the direction of the coronary centerline and radius of the coronary lumen. Leave-one-out cross-validation showed that this allowed accurate and fast tracking of arteries in the eight training CCTA images of the CAT08 evaluation framework.[2] Our experiments showed that the method was able to generalize to images acquired on a different scanner, even though the appearance of coronary arteries in CCTA may differ substantially between scanners [142]. A CNN that was trained using the eight CAT08 training images acquired on a Siemens scanner allowed centerline extraction in 50 CCTA scans acquired on a Philips scanner. Furthermore, we found that in the CAT08 datasets the CNN was able to accurately estimate the radius of vessels within one voxel width, similar to results reported by [143]. This could be used to identify stenoses, to determine artery volumes or as a pre-processing step for accurate coronary lumen segmentation. In the UMCU dataset, there was a small overestimation of vessel radius, which

[2] At the time of writing, the CAT08 evaluation framework was offline due to technical problems. In future work, we will include results on the test set for an independent comparison to other methods.



(a) No augmentation **(b)** Augmentation

Figure 6.15: Sagittal projections of 50 automatically extracted centerlines starting from shifted seed points, for (a) a CNN trained without data augmentation using off-centerline samples, and (b) a CNN trained with off-centerline data augmentation. The original seed point is indicated in red, and manually annotated start point and end point of the vessel are shown in black.

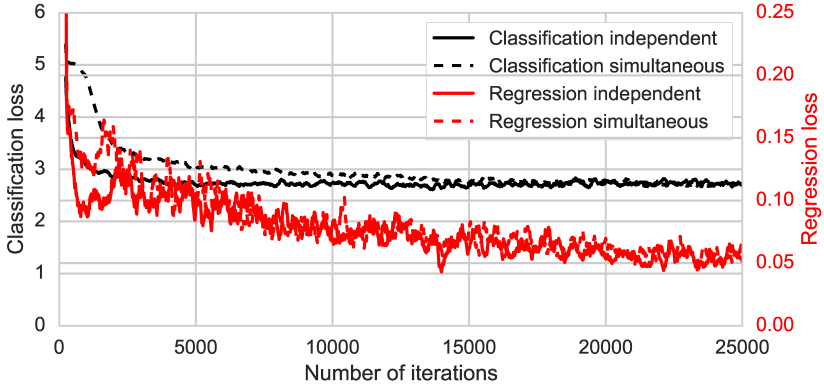


Figure 6.16: Learning curves showing the evolution validation loss for the classification task and the regression task. Results are shown for two CNNs that were trained to perform both tasks independently (solid lines), and one CNN that was trained to perform both tasks simultaneously (dashed lines).

was likely due to the annotation protocol. This will be further investigated in future work.

The method was trained and evaluated with images containing different degrees of calcification, intravascular stents, coronary motion artifacts and step reconstruction artifacts. Whereas previous methods required specific preprocessing steps for e.g. calcium removal [128], we found preprocessing not necessary. Instead, the CNN processed the CCTA data directly, did not require substantial downsampling of images [71] and resulted in a centerline that did not need further refinement, as opposed to [71, 73]. A limitation of the method is that it may occassionally confuse arteries and veins. The CNN only used CCTA HU values, which were not normalized between images. In previous work, we showed that blood pool HU values in CCTA can differ substantially between patients [144]. The attenuation in veins scales accordingly, and therefore veins in one scan may have higher attenuation values than arteries in another scan. In future work, we will investigate pruning methods for extracted centerlines, as proposed in [130].

The presented method is supervised and requires representative training data. Hence, manually annotated reference centerlines are required for a number of training scans. The method was trained with only seven or eight training images, each containing four annotated centerlines. Our experiments showed the largest improvement in validation loss when increasing the training set from two images to four images, and limited improvement when more training scans were included. However, the method would likely benefit from more training data, especially for challenging and rare cases.

The ability of the CNN to generalize to new and unseen data was to a large extent due to augmentation of the training set with rotated samples. This led to substantial invariance of the network to rotations in the input patches. However, we also found that augmentation at test time could further improve the robustness of the method. Such a procedure was previously used by Roth et al. for false positive reduction in lymph node detection [105]. This indicates that the CNN does not obtain full rotation invariance during training. This may be partly due to the discretization of the orientations into classes; the orientation that minimized the CNN's confusion may be found between two discrete orientations. Test time augmentations adds an element of stochasticity, which causes different runs of the algorithm to follow different trajectories [145, 129]. In future work, we will further investigate the robustness of the method.

In this work we have used dilated, or *á trous*, convolution kernels to rapidly increase a CNN's receptive field with few convolution layers, no downsampling layers, and hence no loss of resolution in feature representations [136]. Stacked layers with increasing levels of dilated convolutions have previously successfully extended reception fields in 1D signal processing for audio generation and machine translation [146, 147], and analysis of 2D images [136, 148]. Dilated convolutions are particularly interesting for 3D problems, where the number of parameters typically grows

cubically with the size of the input.

The proposed method is a single-vessel tracking method. In several cases the centerline failed to follow the path to the ostium, and instead followed a side-branch. In the current method, this could be overcome by placement of an extra seed. Alternatively, we found that at branching points the posterior probability distribution over $|D|$ occassionally had not two, but three or four local maxima. These may potentially be used to seed branches, and recover the full coronary tree structure. In future work we will include more training centerlines to investigate the potential of the method for multi-vessel extraction.

6.6 Conclusions

A deep learning-based method for coronary artery centerline extraction has been proposed. The results show that a convolutional neural network can learn to simultaneously determine the direction of coronary artery centerlines and the radius of the coronary lumen.

Chapter 7

Generative adversarial networks for noise reduction in low-dose CT

Based on:
J.M. Wolterink, T. Leiner, M.A. Viergever, I. Išgum "Generative adversarial networks for noise reduction in low-dose CT," *Submitted*

Abstract

Noise artifacts are inherent to low-dose CT acquisition. We propose to train a convolutional neural network (CNN) jointly with an adversarial CNN to estimate routine-dose CT images from low-dose CT images and hence reduce noise.

A generator CNN was trained to transform low-dose CT images into routine-dose CT images using voxel-wise loss minimization. An adversarial discriminator CNN was simultaneously trained to distinguish the output of the generator from routine-dose CT images. The performance of this discriminator was used as an adversarial loss for the generator.

Experiments were performed using CT images of an anthropomorphic phantom containing calcium inserts, as well as patient non-contrast-enhanced cardiac CT images. The phantom and patients were scanned at 20% and 100% routine clinical dose. Three training strategies were compared: the first used only voxel-wise loss, the second combined voxel-wise loss and adversarial loss, and the third used only adversarial loss. The results showed that training with only voxel-wise loss resulted in the highest peak signal-to-noise ratio with respect to reference routine-dose images. However, the CNNs trained with adversarial loss captured image statistics of routine-dose images better. Noise reduction improved quantification of low-density calcified inserts in phantom CT images and allowed coronary calcium scoring in low-dose patient CT images with high noise levels. Testing took less than 10 seconds per CT volume.

CNN-based low-dose CT noise reduction in the image domain is feasible. Training with an adversarial network improves the CNN's ability to generate images with an appearance similar to that of reference routine-dose CT images.

7.1 Introduction

Computed tomography is a widely used imaging modality, allowing visualization of anatomical structures with high spatial and temporal resolution. However, ionizing radiation inherent to CT acquisition continues to raise concerns about potential health hazards [17, 18]. Therefore, the past decade has seen a trend towards dose reduction in CT examinations, and typical dose levels for e.g. coronary CT angiography have been reduced from around 12 mSv in 2009 [149] to 1.5 mSv in 2014 [97].

A drawback of dose reduction is the increase in noise artifacts in reconstructed CT images, which may lead to large local deviations in HU values [150]. A range of methods have been proposed to reduce noise artifacts in low-dose CT, while preserving important details in the image. Iterative reconstruction (IR) techniques have recently been adopted by all major CT vendors [151, 152]. IR techniques iteratively estimate the denoised underlying image and have facilitated high levels of dose reduction [153]. However, these methods require long processing times, dedicated hardware and the availability of projection data.

Besides techniques that operate in the sinogram domain, there is a rich tradition of denoising methods that operate in the image domain, i.e. after image reconstruction from projection data. Nonlocal means filtering methods estimate the noise component based on multiple patches extracted at different locations in the image [154] and have been widely used for CT denoising [155]. Recently, Green et al. proposed a method which replaces noisy image patches by their nearest neighbors in a database of high SNR patches [156]. Alternatively, diffusion filters have been used to sharpen edges and other structures [157].

More recently, several supervised machine learning techniques have been proposed for noise reduction in low-dose CT. Such methods learn a relation between the voxel value in a low-dose image I_{LD} and the voxel value at the same location in a corresponding routine-dose image I_{RD}, based on training with pairs of images. For example, Chen et al. proposed a convolutional neural network (CNN) that estimated routine-dose HU values based on local patches in low-dose CT [158]. This regression method was used to transform low-dose chest and abdomen CT images into estimates of the corresponding routine-dose images. Kang et al. followed a similar approach, but applied the CNN to a contourlet transform of the CT image [159].

The methods proposed in [158] and [159] both showed good quantitative noise reduction properties. However, the parameters of the CNNs were optimized to minimize the per-voxel squared error between the reference routine-dose image and the denoised low-dose image. When estimating a routine-dose CT image, a voxel in the target image may have different possible values, as noise is not only present in the low-dose acquisition, but also in the routine-dose acquisition. Minimizing the squared error between reference and predicted voxel values causes the CNN to predict the mean of these values, resulting in smoothed images that lack the texture

of a typical routine-dose CT image. This smoothing may limit the quantification of small structures in denoised images.

In this work, to overcome the limitations of voxel-wise regression in noise reduction, we propose to train a noise reducing *generator* CNN together with an adversarial *discriminator* CNN, as a generative adversarial network (GAN) [160, 161]. The discriminator CNN aims to differentiate between real routine-dose CT images and transformed low-dose CT images. The performance of this discriminator CNN adds a loss term to optimization of the generator CNN, forcing the generator to provide more realistic estimates of the routine-dose CT image. The discriminator CNN is only used to provide feedback during training, and thus adds no complexity at test time.

In addition, we address the limitation that spatially aligned low-dose and routine-dose CT images are required for training of a noise reducing CNN. Well-aligned clinical scans acquired at different dose levels are often not available. Therefore, previous works have resorted to simulation of low-dose CT images based on routine-dose images [156, 158], which is a challenging problem [162]. Here, we show that a generator CNN trained with only adversarial feedback can learn the appearance of routine-dose CT images, without spatially aligned low-dose and routine-dose images.

The method is demonstrated on CT scans of an anthropomorphic phantom acquired at low dose and routine clinical dose and non-contrast-enhanced cardiac CT scans of patients who were scanned with a low dose and a routine clinical dose. We quantitatively analyze the noise in filtered back projection (FBP) reconstructed images at these two dose levels, as well as in low-dose FBP images transformed by our method, and show that the method has strong noise reducing properties while preserving the texture in the CT scan.

Cardiac CT images without contrast enhancement are clinically used to quantify coronary artery calcification (CAC), which is a strong and independent predictor of cardiovascular events [8]. For CAC quantification, connected components in the coronary arteries above 130 HU are identified. However, excessive noise levels in low-dose CT scans make it difficult if not impossible to limit quantification to CAC lesions only. This may cause large overestimations of CAC or render images non-interpretable [163]. While noise reduction offers a solution to this problem, strong noise reduction may negatively affect the quantification or identification of small and low-density CAC lesions [164, 165]. We show how the proposed use of an adversarial network improves CAC quantification over standard FBP at low-dose and over a generator trained without an adversarial network. Furthermore, we show that the proposed method allows CAC quantification in low-dose patient images.

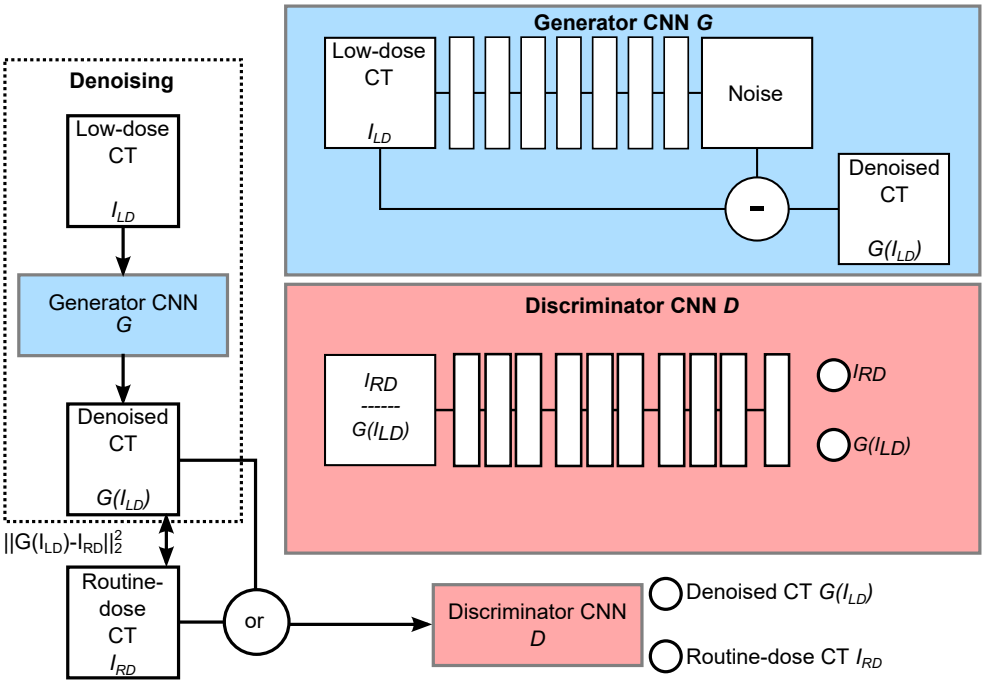


Figure 7.1: Overview of the proposed pipeline for noise reduction in low-dose CT. The generative adversarial network consists of two components: a generator CNN and a discriminator CNN. The generator uses regression to determine the routine-dose HU value at every voxel in a low-dose CT. It does this through a skip connection which subtracts an estimated noise image from the input low-dose image. The discriminator tries to distinguish reduced noise CT images from real routine-dose images.

7.2 Methods

Fig. 7.1 illustrates the proposed system, which has two parts. The first part consists of a generator CNN G that analyzes a low-dose CT image I_{LD}. The generator returns $G(I_{LD})$, which is an approximation of a routine-dose CT image I_{RD}. The system has two ways to enforce similarity between $G(I_{LD})$ and I_{RD}. First, if voxels in I_{LD} and I_{RD} are spatially aligned, the voxel-wise error between $G(I_{LD})$ and I_{RD} may be minimized. Second, a discriminator CNN D may be simultaneously trained to differentiate between $G(I_{LD})$ and I_{RD}. If the discriminator can make this distinction easily, i.e. if the generated CT images do not resemble real routine-dose images, the generator needs to improve its estimations.

Hence, both networks have different tasks. While the generator performs regression of voxel values, the discriminator performs classification of images.

7.2.1 Generator CNN

The generator CNN G transforms the low-dose CT image I_{LD} into an image with a reduced noise level $G(I_{LD})$ approximating the reference routine-dose image I_{RD}. We assume that $I_{LD} = I_{RD} + N$, i.e. the noisy image is the reference routine-dose image with superimposed noise N. Hence, the task of the trainable layers in the CNN can be simplified to prediction of the noise N.

Cardiac CT images are typically anisotropic, with larger voxel spacing in the cranio-caudal direction. Therefore, the input to the generator CNN was a 3D rectangular volume of $65 \times 65 \times 19$ voxels. The network contains seven consecutive convolution layers with small convolution kernels of size $3 \times 3 \times 3$ voxels [108]. The number of kernels increases from 32 in the first layers, to 64 and 128 in the final layers. No padding is applied after convolutions. Hence, a receptive field of $15 \times 15 \times 15$ voxels in the input determines the result for a voxel in the output, which has size $51 \times 51 \times 5$ voxels. The final layer outputs the predicted noise through a linear activation function. The noise is then subtracted from the low-dose CT image to return a noise-reduced image $G(I_{LD})$.

All trainable layers except the final layer use leaky rectified linear activation functions (LReLUs) to increase training stability [166, 161]. Because the noise values are relatively small compared to the range of possible CT HU values, we initialize the weights in the convolution layers to a normal distribution ($\mu = 0.0, \sigma = 0.001$). Batch normalization [137] is used in the generator CNN, but not directly after the input layer of directly before the output layer.

7.2.2 Discriminator CNN

The discriminator takes either a routine-dose CT subimage I_{RD} or a processed low-dose CT subimage $G(I_{LD})$ as input, and determines whether the input is a real routine-dose image or not.

The input to the discriminator is a 3D rectangular volume of $51 \times 51 \times 5$ voxels, which is the size of the generator's output. The first two convolution layers use $3 \times 3 \times 3$ convolution kernels, which reduce the size of the volume to $47 \times 47 \times 1$ voxels. Hence, subsequent layers operate on 2D feature maps and consequently use 3×3 convolution kernels. Convolution layers are organized in three blocks. Each block contains two layers without strided convolution, and one layer with a stride of 2 [167]. The resulting feature maps are connected to an output node through a hidden layer with 256 nodes. As in the generator, we use LReLu activation functions and batch normalization. The final layer contains a sigmoid activation to determine whether the input is a real routine-dose CT image (label 1) or not (label 0). Weights in the discriminator network are initialized using the method proposed in [112].

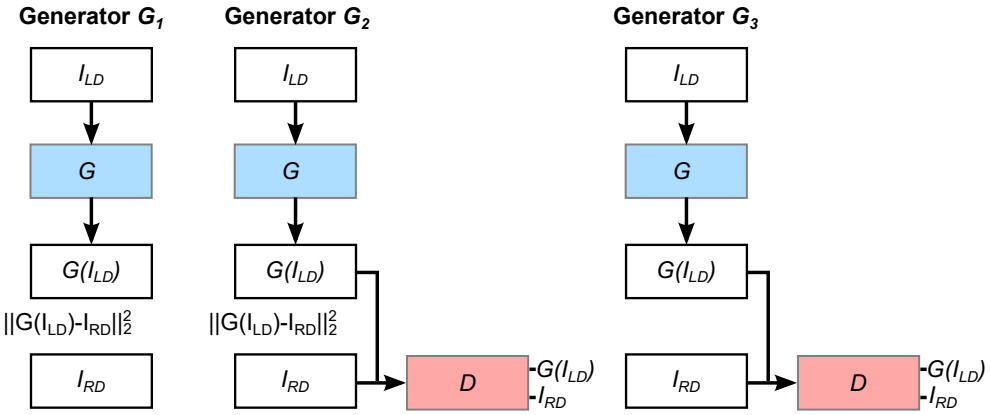


Figure 7.2: Three different approaches to CNN-based low-dose CT noise reduction. Generator G_1 is trained only with the squared error loss between the reduced noise image and the routine-dose image. Generator G_2 is trained with the squared error and the discriminator feedback. Generator G_3 is trained with only the discriminator feedback.

7.2.3 Training

The system in Fig. 7.1 contains output $G(I_{LD})$ of the generator and outputs $D(G(I_{LD}))$ or $D(I_{RD})$ of the discriminator. Two loss components directly affect the generator. First, the squared error between $G(I_{LD})$ and I_{RD}. Second, the ability of the discriminator to identify images generated by the generator. Hence, the loss for the generator is defined as

$$\ell_G = \lambda_1 \|G(I_{LD}) - I_{RD}\|_2^2 + \lambda_2 \ell_{bce}(D(G(I_{LD})), 1), \tag{7.1}$$

where $\ell_{bce}(D(G(I_{LD})), 1)$ is the binary cross-entropy between the discriminator's prediction and the label 1 that the generator wants the discriminator to predict. The parameters λ_1, λ_2 determine a weighting between the regression error of the generator and the classification error of the discriminator.

The discriminator has an adversarial goal to the generator and wants to correctly distinguish the processed low-dose CT images from the routine-dose CT images. Hence, the discriminator minimizes

$$\ell_D = \ell_{bce}(D(I_{RD}), 1) + \ell_{bce}(D(G(I_{LD})), 0), \tag{7.2}$$

where $\ell_{bce}(D(I_{RD}), 1)$ is the binary cross-entropy between the discriminator's decision on routine-dose CT samples and their target label 1, and $\ell_{bce}(D(G(I_{LD})), 0)$ is the binary cross-entropy between the discriminator's decision on processed low-dose CT samples and their target label 0.

The loss of the generator (Eq. 7.1) can be optimized in three ways. First, by only incorporating the voxel-wise loss between $G(I_{LD})$ and I_{RD}, i.e. $\lambda_1 > 0, \lambda_2 = 0$, as was previously proposed in [158, 159]. Second, by combining the voxel-wise

loss with the loss incurred through the discriminator ($\lambda_1 > 0, \lambda_2 > 0$). Third, by optimization based only on the discriminator feedback ($\lambda_1 = 0, \lambda_2 > 0$). In this case, an L2 regularization term on the noise map is added to the loss function. For the first two strategies, voxel-alignment between I_{LD} and I_{RD} is required, while no such alignment is required for the third strategy. We name the resulting trained generator CNNs G_1, G_2 and G_3 (Fig. 7.2).

All relevant parameters in the generator and discriminator are simultaneously optimized using the Adam optimizer [138], with a learning rate of 0.0002 and an exponential decay rate for the first moment estimates 0.5 [161]. While related methods on GAN-based image transformation proposed to alternate between optimization of the generator and the discriminator during training [168, 169], we found that simultaneously optimizing both networks led to more stable behavior of the system. The method was implemented in Theano and Lasagne. Experiments were run on a NVIDIA Titan X GPU with 12 GB memory.

7.3 Data

We include low-dose and routine-dose CT scans of an anthropomorphic thorax phantom, as well as low-dose and routine-dose patient cardiac CT scans.

7.3.1 Phantom CT scans

An anthropomorphic thorax phantom (QRM anthropomorphic thorax phantom; QRM GmbH; Möhrendorf; Germany) with a central recess was used. This recess was filled with water, in which one of two artificial coronary arteries was placed. The first coronary artery contained two inserts with densities of 196 and 380 mg hydroxyapatite (HA)/cm^3, the second coronary artery contained two inserts with densities of 408 and 800 mg HA/cm^3. All inserts had the same dimensions, and a volume of 196.3 mm^3. The phantom was embedded in an extension ring to simulate CT acquisition of average-sized patients.

The phantom was scanned on a Philips Brilliance iCT 256 scanner (Philips Healthcare, Best, The Netherlands), with a tube voltage of 120 kVp. Images were acquired in sequential mode at 20% routine dose (10 mAs) and 100% routine dose (50 mAs) [170]. The phantom was scanned at these two dose levels, so that low-dose and routine-dose images were spatially aligned. After this, a small translation and rotation were applied to the phantom. This process was repeated five times. Scanning was performed with the artificial artery containing 196 and 380 mg HA/cm^3 inserts and with the artificial artery containing 408 and 800 mg HA/cm^3 inserts. Hence, the image set contained five acquisitions per dose level per artificial artery. Images were reconstructed using filtered backprojection (FBP) to a slice thickness and increment of 3.0 mm, matching the parameters for cardiac CT reconstruction. In-plane resolution was 0.49 mm.

7.3.2 Cardiac CT scans

Non-contrast-enhanced cardiac CT scans of 28 patients were previously obtained in a study which was approved by the local institutional review board and for which written informed consent from all participants was obtained [171]. Patients were scanned on a Philips Brilliance iCT 256 scanner (Philips Healthcare, Best, The Netherlands), using a tube voltage of 120 kVp and a tube current of either 10/50 mAs or 12/60 mAs, depending on the patient's weight (threshold $\geq$ 80 kg). Hence, two scans were acquired: one at 20% and another at 100% routine cardiac dose, with an effective dose of 0.2 or 0.9 mSv. Image acquisition was ECG-triggered. All images were reconstructed using FBP with slice thickness 3.0 mm and slice increment 1.5 mm. In-plane resolution ranged from 0.35 to 0.49 mm.

7.3.3 Evaluation

Noise levels were characterized using the mean and standard deviation of HU values in a homogeneous region of interest. The Friedman test was used to analyze noise differences among reconstructions and the Wilcoxon signed-rank test with Bonferroni correction was used to analyze pairwise differences.

The correspondence between voxel-aligned images was quantitatively evaluated using the peak signal-to-noise ratio (PSNR). In the case of CT, the PSNR is defined as

$$\mathrm{PSNR} = 20 \log_{10} \frac{4095}{\sqrt{MSE}}, \tag{7.3}$$

where 4095 is the maximum range of HU values, and MSE is the mean squared error between two images. In all cases, the routine-dose FBP image was used as the reference standard.

The effect of noise reduction on the quantification of coronary artery calcification was analyzed. In the phantom CT images, we quantified the volume (in mm^3) and mass (in mg) of the four inserts. In the patient scans, we quantified total calcium burden using the Agatston score, which is a clinically used intensity-weighted measure of calcified area [9].

7.4 Experiments and results

7.4.1 Noise reduction

To investigate the ability of the method to reduce noise, we performed experiments using synthesized 1D signals, as well as the 3D phantom and patient cardiac CT images described in the previous section.

Synthetic 1D signals

To illustrate the characteristics and differences among generators G_1, G_2 and G_3, we performed experiments with 1D signals. The architecture of the generator and discriminator CNN was the same as described in the previous section, but all 2D and 3D operations were replaced by 1D operations.

Training samples with length 119 were randomly generated and initialized to 0. A block activation with a random length and value between 0 and 500 HU was added at a random location in the signal. To synthesize low-dose and routine-dose signals for generators G_1 and G_2, Gaussian noise with $\mu = 0$ and $\sigma = 70$ or $\sigma = 20$ was added to the base signal. This is similar to noise levels observed in the CT scans. To train generator G_3, separate base signals were acquired for the low-dose and routine-dose signal. All generator were trained for 5,000 iterations with mini-batches of 48 samples.

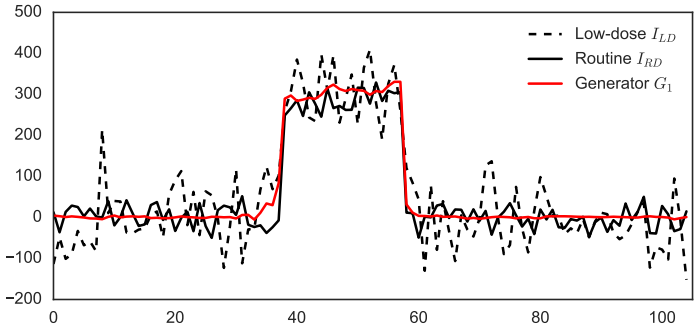
Fig. 7.3 shows the synthetic low-dose signal I_{LD}, the target synthetic routine-dose signal I_{RD} and the low-dose signal signal after noise reduction, i.e. $G_1(I_{LD})$, $G_2(I_{LD})$ and $G_3(I_{LD})$. Signal $G_1(I_{LD})$ shows that the generator has learned to smooth the signal, and to predict values with low standard deviation. In contrast, generators G_2 and G_3 have both learned to reduce the noise level in the signal to that of the routine-dose signal I_{RD}.

Phantom CT scans

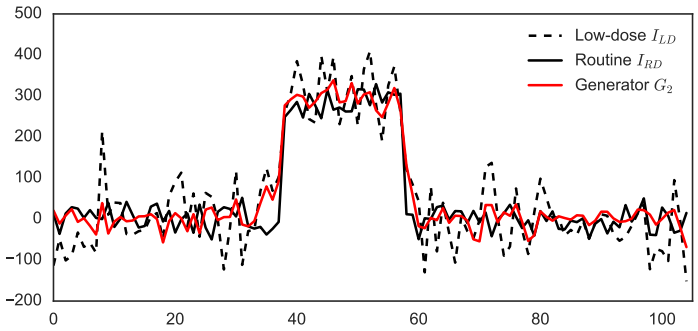
Voxels in the phantom scans were aligned between images acquired at different dose levels. Hence, voxel-wise loss could be used during training. Generators G_1, G_2 and G_3 were each trained using five-fold cross-validation. Each fold contained one scan with 196 and 380 mg HA/cm^3 inserts, and one scan with 408 and 800 mg HA/cm^3 inserts. The generators were trained to transform 10 mAs low-dose CT images into 50 mAs routine-dose CT images. Each network was trained for 5,000 iterations, with mini-batches of 48 samples.

Fig. 7.4 shows axial images of FBP, G_1, G_2 and G_3 reconstructions of the phantom scanned at low-dose, as well as an FBP reconstruction of the phantom scanned at routine-dose. It can be seen that the low-dose FBP reconstruction in Fig. 7.4a shows substantial noise artifacts, with local deviations up to 70 HU. Generator G_1 (Fig. 7.4b) generates a smooth image with low noise levels. Generators G_2 (Fig. 7.4c) and G_3 (Fig. 7.4d) generate a slightly noisier image with a noise profile that better matches that of the reference image (Fig. 7.4e). Note that in the images generated by G_1, G_2 and G_3 a ring artifact remains visible after noise reduction. This artifact has a larger scale than the receptive field of the generator and discriminator and are thus not removed.

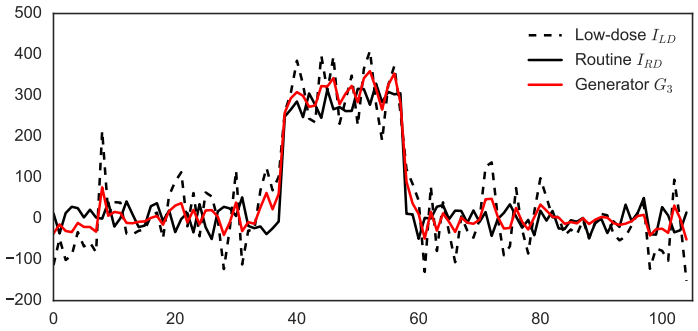
During training, reconstructions for G_1, G_2 and G_3 were obtained after every 500 iterations. In each reconstruction, the noise level was determined in a homogeneous $32 \times 32 \times 2$ voxel ROI in the central recess (red squares in Fig. 7.4). Fig. 7.5a shows



(a) Generator G_1

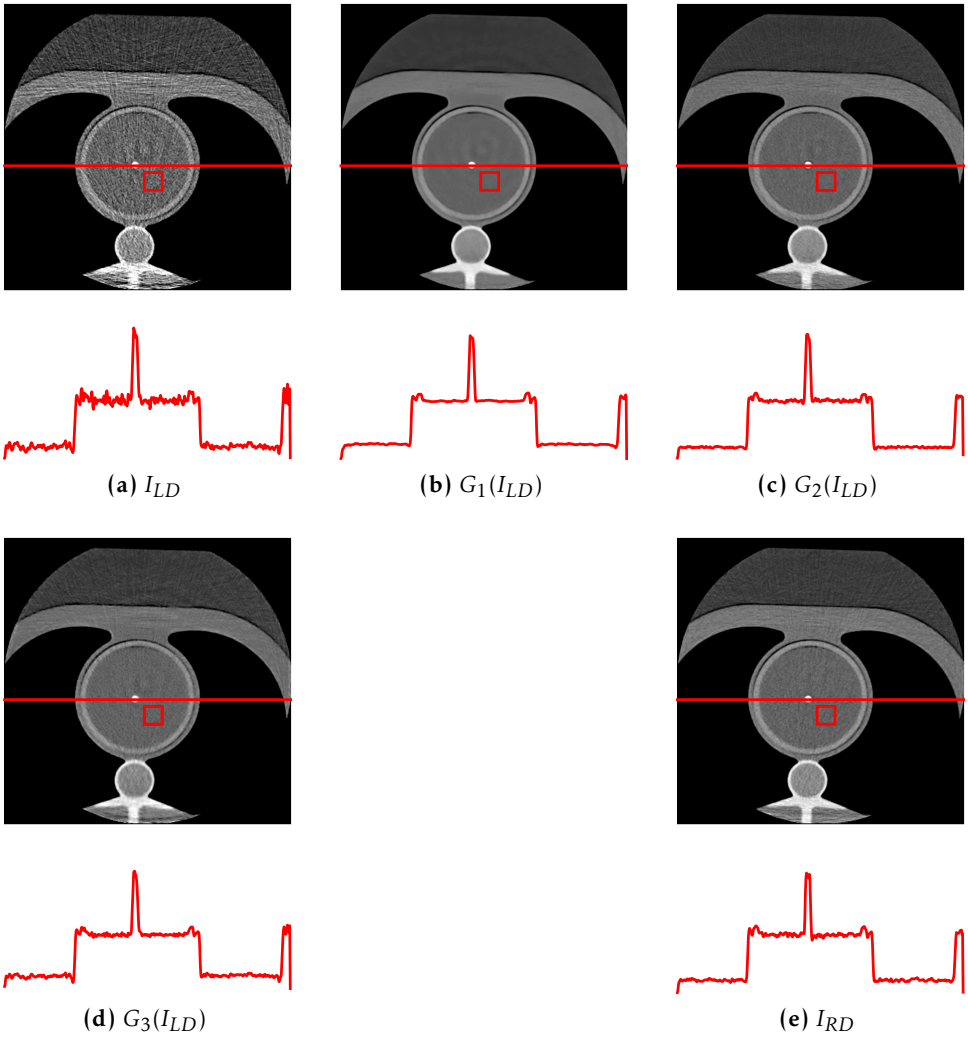


(b) Generator G_2



(c) Generator G_3

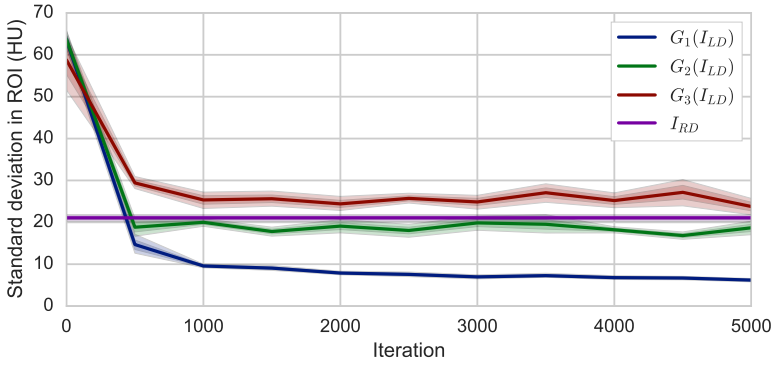
Figure 7.3: Reducing noise in a synthesized 1D low-dose signal I_{LD} with generators G_1, G_2 and G_3. Generator G_1, which was trained without an adversarial discriminator, predicts smooth signals with low standard deviation, which do not resemble the synthetic routine-dose signal I_{RD}. The generators that were trained with an adversarial discriminator predict noisy values in the same distribution as the synthetic routine-dose signal I_{RD}.



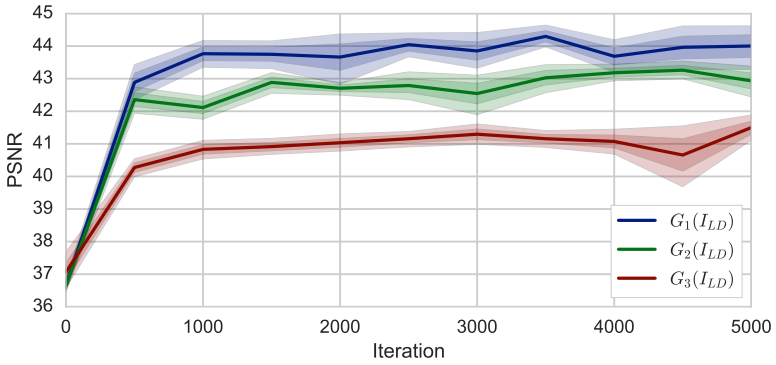
(a) I_{LD} **(b)** $G_1(I_{LD})$ **(c)** $G_2(I_{LD})$

(d) $G_3(I_{LD})$ **(e)** I_{RD}

Figure 7.4: Cardiac CT phantom with insert. (a) FBP-reconstructed low-dose CT image I_{LD} acquired at 10 mAs, (b) low-dose CT image processed by generator G_1, (c) low-dose CT image processed by generator G_2, (d) low-dose CT image processed by generator G_3, (e) reference routine-dose CT I_{RD} acquired at 50 mAs. Red lines in the images indicate intensity profiles, shown below the image. Red squares indicate ROIs used for noise measurement.



(a) Standard deviation in ROI (HU)



(b) PSNR

Figure 7.5: (a) Standard deviation in a homogeneous ROI in the phantom CT scans, for generators G_1, G_2 and G_3 during generator training. The line for I_{RD} indicates the noise level in routine-dose FBP reconstructions. (b) Peak-signal-to-noise ratio (PSNR) between generated images and routine-dose FBP during generator training. In both cases the average of five-fold cross-validation is shown.

the evolution of the standard deviation in this ROI during training, where a reference line indicates the noise distribution in the routine-dose FBP images I_{RD}. For all generators, the noise level is initially the same as in the low-dose CT scan. While the standard deviation in scans reconstructed by G_1 continues to decrease throughout training, the standard deviation in scans reconstructed by G_2 plateaus around the level of the routine-dose scan. For G_3, the noise plateaus above the level of the routine-dose scan, indicating a less accurate estimation of I_{RD}. Fig. 7.5b shows the average PSNR during training. The PSNR is highest for scans generated by generator G_1, which minimizes the squared error between the low-dose and routine-dose image.

Cardiac CT scans

In clinically acquired cardiac CT, we may not expect perfect voxel-wise correspondence between scans acquired at different dose levels, due to motion and breathing of the patient. Image registration may mitigate this problem to some extent, but also adds unwanted transformations to the image and noise patterns. Therefore, generator G_3 was trained using only adversarial loss.

The set of 28 patient scans was separated into two sets of 14 patients for a two-fold cross-validation, where patients were once in the training set and once in the test set. The generator was trained to transform low-dose images into routine-dose images. Training was performed as in the phantom scans, for 5,000 iterations with mini-batches of 48 3D samples.

To quantify the noise in each reconstruction, an ROI was placed in the ascending aorta at the level of the left coronary ostium. The mean HU values in this ROI showed no statistically significant difference between the different reconstructions (Fig. 7.6). However, there was a significant difference among the standard deviations in these ROIs ($p<0.001$), with a median (IQR) value of 60.5 (55.8–77.9) HU for I_{LD}, 28.7 (25.2–35.1) for $G_3(I_{LD})$ and 25.6 (22.9–30.4) for I_{RD}. Post-hoc testing with the Wilcoxon signed-rank test revealed significant differences ($p<0.001$) between I_{LD} and I_{RD} and I_{LD} and $G_3(I_{LD})$. Hence, application of generator G_3 resulted in significant noise reduction.

7.4.2 Coronary calcium quantification

Phantom CT scans

In each phantom reconstruction, the volume and mass of two inserts were quantified. The inserts were identified and segmented using a clinically used threshold of 130 HU and 26-connectivity region growing [9]. Fig. 7.7 shows the volume and mass of each of the four inserts in different reconstructions. The reference volume of each insert was 196.3 mm^3, indicated by the red line. In the low-dose FBP reconstruction, the volume and mass of the inserts were systematically overestimated due to noise $\geq$ 130 HU in the image. In one acquisition, both inserts were connected to

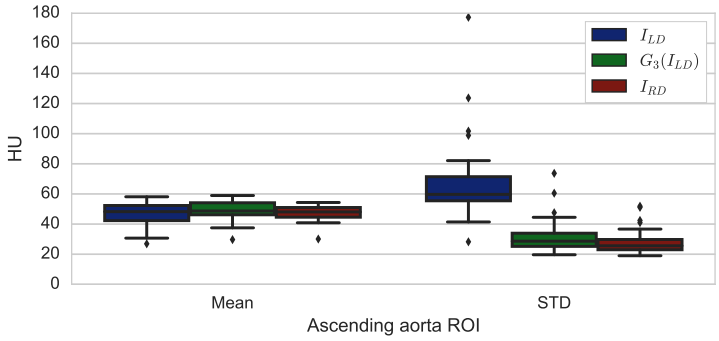
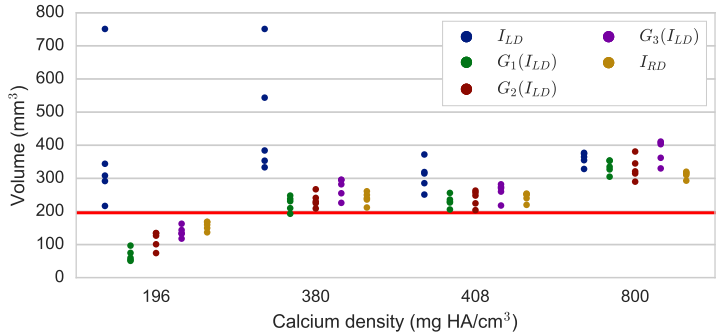
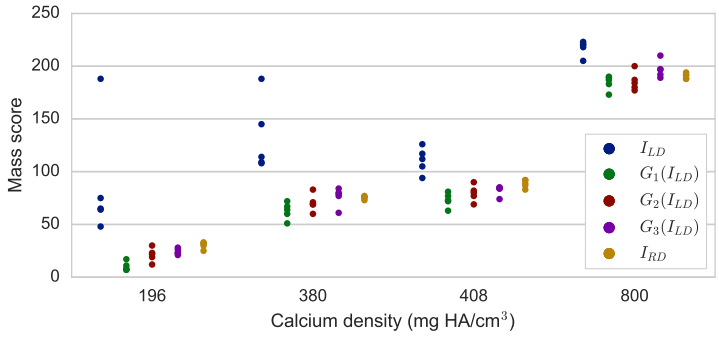


Figure 7.6: Mean and standard deviation in an ROI in the ascending aorta at the level of the left coronary ostium. While the mean HU values do not change between low-dose FBP reconstruction, reconstruction with generator G_3 and routine-dose FBP reconstruction, there is a statistically significant difference between the standard deviation in low-dose FBP and the other two reconstructions.



(a) Volume



(b) Mass

Figure 7.7: (a) Volume and (b) mass quantification of coronary calcification inserts in low-dose (I_{LD}) and routine-dose (I_{RD}) FBP images of the phantom, as well as images reconstructed using generators G_1, G_2 and G_3.

each other by voxels ≥ 130 HU. All three generators reduced noise in a way that led to more accurate quantification of calcium volume. However, generator G_1 trained using only voxel-wise square error led to the largest underestimation of the volume of the insert with the lowest density. Similarly, the mass score was most underestimated when G_1 was used.

Cardiac CT scans

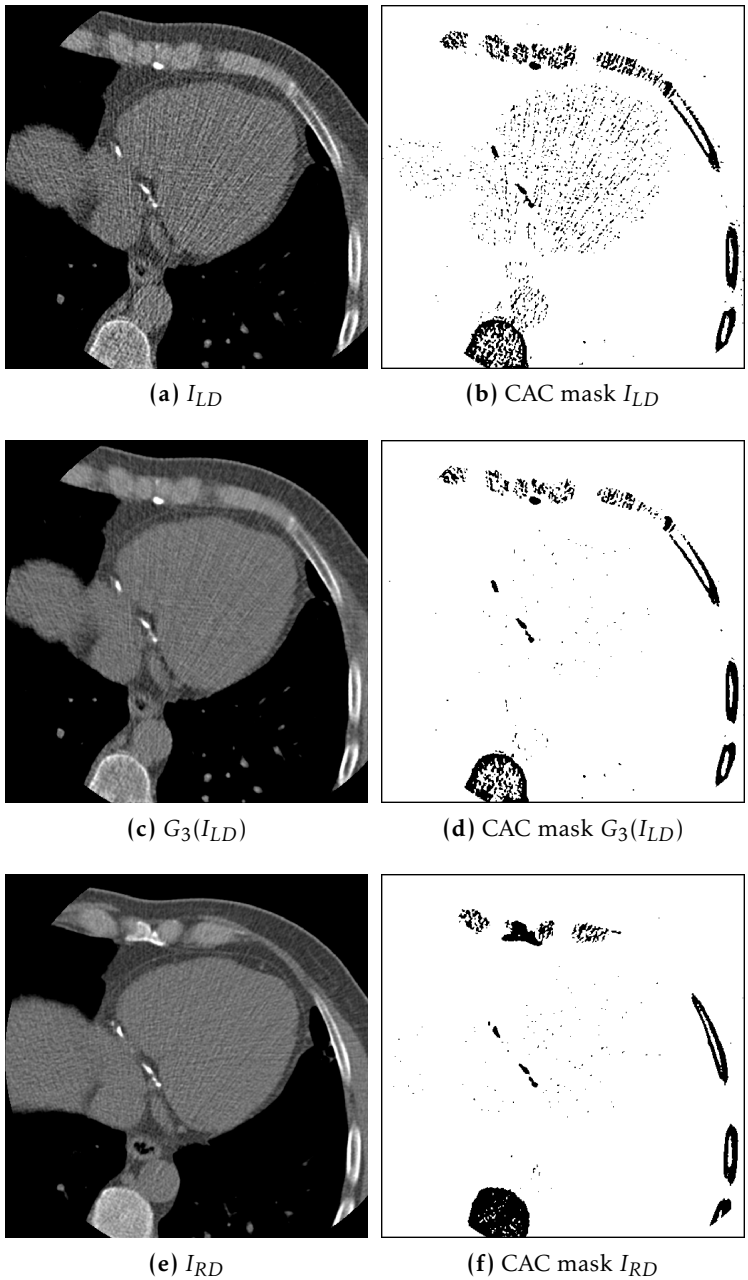
Fig. 7.8 shows a low-dose cardiac CT FBP image, the same image transformed by generator G_3 and the routine-dose FBP image of the same patient. For each image, a CAC candidate mask is shown, indicating all voxels ≥ 130 HU. In the low-dose FBP image, calcium scoring is difficult if not impossible due to large clusters of connected voxels with density ≥ 130 HU. However, after noise reduction with generator G_3, the distribution of noise voxels is similar to that in the routine-dose scan, and the CAC lesions remain above the density threshold. Note the difference between the anatomy visualized in the low-dose scan and the routine-dose scan, making voxel-wise alignment difficult.

Coronary calcium was quantified in the low-dose images, the images generated by generator G_3, and the routine-dose images. In 9/28 low-dose images, noise levels were too high to perform reliable calcium scoring. After noise reduction with generator G_3, noise levels were too high in 3/28 images. These were the scans with the highest noise levels in low-dose CT FBP, with standard deviations in the aortic ROI of 177.3, 144.5 and 123.7 HU, respectively. Calcium could reliably be scored in all routine-dose images. Out of 19 patients for whom calcium could be scored in I_{LD}, $G_3(I_{LD})$ and I_{RD}, 11 contained coronary calcium. In these images, median (IQR) Agatston scores were 72.7 (19.8–204.5) in I_{LD}, 25.9 (4.9–145.1) in $G_3(I_{LD})$, and 53.5 (21.1–183.0) in I_{RD}.

7.5 Discussion

We have described a method to reduce noise in low-dose CT images using convolutional neural networks. The results show that training with adversarial feedback from a discriminator CNN can generate images with a more similar appearance to the routine-dose CT than training without a discriminator CNN. Feedback from the discriminator prevents smoothing in the image and allows more accurate quantification of low-density calcifications in phantom CT scans.

The results show that the proposed method is capable of substantial noise reduction in phantom CT images, and that combining a voxel-wise squared error loss with adversarial loss led to a noise distribution that was similar to that in the reference routine-dose image. Training with only squared error loss as proposed in [158] led to smooth images with low noise levels, while training with only adversarial loss led to images with slightly higher noise levels than in the routine dose. Hence, in practice it would be good to combine the two loss components when possible.



(a) I_{LD} **(b)** CAC mask I_{LD}

(c) $G_3(I_{LD})$ **(d)** CAC mask $G_3(I_{LD})$

(e) I_{RD} **(f)** CAC mask I_{RD}

Figure 7.8: Example CT slice for (a) 20% dose FBP reconstruction I_{LD} and (b) corresponding artery calcification (CAC) scoring mask, (c) 20% dose generator G_3 reconstruction $G_3(I_{LD})$ and (d) corresponding CAC scoring mask, and (e) routine-dose FBP reconstruction I_{RD} and (f) corresponding CAC scoring mask.

In cardiac CT images, where spatially aligned training images are not available, we found that a generator trained with only adversarial loss was able to significantly reduce noise levels, while preserving mean tissue HU values. This shows that a convolutional neural network can learn to reduce noise in low-dose CT images even when no spatially aligned routine-dose scan is available. In previous studies, the problem of voxel alignment was mitigated by simulation of low-dose CT images based on routine-dose images [156, 158]. However, realistic low-dose CT simulation is a challenging problem, and simulated scans do not necessarily resemble real low-dose acquisitions [162]. In future work, we will investigate whether generator G_3 can also learn to reduce noise when the low-dose and routine-dose sets consist of different patients, i.e. if inter-patient learning is possible.

The phantom calcium quantification results indicate that the generator trained using only squared error applies too much smoothing to the image, and affects quantification of low-density calcifications. Training with adversarial feedback resulted in calcium scores that were closer to those obtained in the routine-dose image. In all cases, calcium scores in the noise-reduced images were lower than those in the low-dose FBP reconstructions, which is also a common observation in clinically used sinogram-based IR noise reduction [164, 165]. In patient cardiac CT scans, noise reduction allowed calcium scoring in six scans that previously contained too much noise. There were 11 patiens with coronary calcium whose images allowed scoring in I_{LD}, $G_3(I_{LD})$ and I_{RD}. For these patients, the obtained calcium scores in reduced noise images were lower than those in low-dose CT, which corresponds to our observation in the phantom images. Similarly, calcium scores obtained in $G_3(I_{LD})$ were generally lower than those in routine dose scans. This could be due to noise reduction, but may also be caused by interscan variability [172], and should be further investigated in a larger sample.

Smoothing effects in CNN-based image-to-image regression have been addressed in computer vision as well as medical image analysis. For colorization of grayscale natural images, Zhang et al. proposed to pose the regression problem as a classification problem with one class for each potential pixel value. In our problem this would mean prediction of 4096 highly unbalanced classes, which is infeasible. Alternatively, GANs have been used for a wide range of image-to-image transformations, including image colorization, super-resolution [173], video frame prediction [174], segmentation [168] or general image-to-image conditioning [175]. In medical image analysis, Nie et al. used a GAN to transform MRI images into estimates of CT images [169].

While methods for semantic image segmentation typically require CNNs with large receptive fields [27, 148], noise is localized. In the current approach, the generator has a relatively small receptive field of $15 \times 15 \times 15$ voxels. This is in line with patch-based methods such as presented in [156] and the finding of [150] that such receptive fields are sufficient to estimate the local extent of noise. In contrast, streak artifact removal may require CNNs with larger receptive fields [176]. The

proposed generator CNN does not directly predict HU values in the routine-dose CT image, but predicts the amount of additive noise at each position in the image. An additional operation is performed to subtract this noise from the image. A similar approach was recently proposed by Han et al. for subtraction of streak artifacts from CT images reconstructed with a low number of projections [176]. Previously proposed GAN methods also used the input to G as input to D [168, 175]. We found that this led to a strong bias towards the discriminator's performance and was infeasible in the case of generator G_3.

In the proposed method, the discriminator network performs an auxiliary task during training and the network is not used during testing. However, after training, the discriminator has learned to extract certain features from low-dose and routine-dose non-contrast-enhanced cardiac CT images. In future work we will investigate if these features could be useful for other tasks in non-contrast-enhanced cardiac CT, such as automatic coronary calcium scoring [15, 16, 33].

A potential limitation of the current method is that pathologies might be introduced that are actually not present in the image, based on their presence in the training set. In future work, it would be interesting to estimate the certainty of the method at each location in the image. A major advantage of the proposed method is its processing speed. The discriminator CNN is only used during training, which restricts computational requirements during testing. The method has a runtime of less than 10 s on a $512 \times 512 \times 90$ CT volume, and may thus be efficiently applied to an already reconstructed low-dose scan, without the need for sinogram data.

7.6 Conclusion

Low-dose CT noise reduction in the image domain using a convolutional neural network is feasible. Training with an adversarial network allows the generator to better learn the noise distribution in routine-dose CT and produce more realistic images for more accurate coronary calcium quantification.

Chapter 8

Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease

Based on:
J.M. Wolterink, T. Leiner, M.A. Viergever, I. Išgum "Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease," *MICCAI 2016 HVSMR Workshop*, 2017, Lecture Notes in Computer Science, vol. 10129, pp. 95–102

Abstract

We propose a method using dilated convolutional neural networks (CNNs) for automatic segmentation of the myocardium and blood pool in cardiovascular MR (CMR) of patients with congenital heart disease (CHD).

Ten training and ten test CMR scans cropped to an ROI around the heart were provided in the MICCAI 2016 HVSMR challenge. A dilated CNN with a receptive field of 131×131 voxels was trained for myocardium and blood pool segmentation in axial, sagittal and coronal image slices. Performance was evaluated within the HVSMR challenge.

Automatic segmentation of the test scans resulted in Dice indices of 0.80 ± 0.06 and 0.93 ± 0.02, average distances to boundaries of 0.96 ± 0.31 and 0.89 ± 0.24 mm, and Hausdorff distances of 6.13 ± 3.76 and 7.07 ± 3.01 mm for the myocardium and blood pool, respectively. Segmentation took 41.5 ± 14.7 s per scan.

In conclusion, dilated CNNs trained on a small set of CMR images of CHD patients showing large anatomical variability provide accurate myocardium and blood pool segmentations.

8.1 Introduction

Congenital heart diseases (CHD) are a type of congenital defect affecting almost 1% of live births [21]. Patients with severe congenital heart disease often require surgery in their childhood. It has been shown that the use of patient-specific 3D models is helpful for preoperative planning [22]. Such models are typically based on a segmentation of the patient's anatomy in cardiovascular MR (CMR). However, segmentation of cardiac structures in CMR requires several hours of manual annotations [23]. Hence, there is a need for semi-automatic or fully automatic segmentation methods to speed up this time-consuming process and reduce the workload for clinicians.

The large anatomical variability among patients poses a major challenge for (semi)automatic segmentation of CMR in CHD patients (Fig. 8.1). Methods relying on multi-atlas based segmentation would require a highly diverse training set representing the various manifestations of CHD. Hence, local analysis based on intensity and texture might be advantageous. Pace et al. proposed a patch-based semi-automatic segmentation method that produces highly accurate segmentations, requiring one hour of manual interaction and one hour of offline processing per scan [177]. The label of each voxel in an image is determined based on patch similarity to manually segmented sections in the image. To eliminate any user interaction, we propose a fully automatic patch-based voxel classification method. Voxel labels are determined based on similarities to voxels in training images using a convolutional neural network (CNN).

CNNs have been widely adopted in medical image analysis for segmentation of e.g. tissue classes [27] and tumors [178] in brain MR, neuronal structures in electron-miscroscopy [179] and coronary artery calcium in cardiac CT angiography [144]. A CNN labels each voxel in an image based on one or multiple patches surrounding that voxel. An effective voxel classification method should combine both local structure information and global context information. To this end, Moeskops et al. proposed a multi-scale approach using differently-sized patches extracted for every voxel [27], and Ronneberger et al. used a CNN which merges information at different scales by skipping layers [179]. However, patch extraction at every voxel is time-consuming and downsampling layers affect the output resolution and translational equivariance, meaning that the exact output may depend on the positioning of the input.

Recently, stacks of dilated convolutions have been proposed for image segmentation [136]. Such stacks aggregate features at multiple scales through convolutional layers with very few parameters, thereby avoiding problems such as overfitting, while generating high resolution output images with translation equivariance. The promise of a large receptive field with few trainable parameters is particularly interesting in medical imaging, where data sets are often small. In this work, we use CNNs with dilated convolutions to automatically segment CMR images of CHD patients.

8.2 Data

The method was developed and evaluated within the framework of the MICCAI Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease (HVSMR 2016)[1]. Ten training and ten test CMR scans were provided by the workshop organizers. The scans were acquired at Boston Children's Hospital with a 1.5T Philips Achieva scanner (TR = 3.4 ms, TE = 1.7 ms, $\alpha = 60°$). Three images were provided for each patient: a complete axial CMR image, the same image cropped around the heart and thoracic aorta, and a cropped short axis reconstruction. In the current work, we use the cropped image around the heart and thoracic aorta. Reconstructed in-plane resolution of the images ranged from 0.73 mm to 1.15 mm, while slice spacing ranged from 0.65 mm to 1.15 mm.

Reference segmentations of the blood pool and myocardium were provided for the training scans, but not for the test scans. These segmentations were made by a trained observer and validated by two clinical experts. The blood pool class contained both atria and ventricles, the aorta, pulmonary veins, and superior and inferior vena cava. The myocardium class contained the thick muscle around the two ventricles and their separating septum. Example reference segmentations are shown in Fig. 8.1.

8.3 Methods

We trained a purely convolutional CNN to assign a class label to each voxel in a CMR volume based on classification of three orthogonal patches centered at the voxel. The CNN uses dilated convolutions allowing large receptive fields with few trainable parameters.

CNNs consist of a sequence of convolution layers, which convolve an image F_l at layer l with a kernel k to obtain image F_{l+1} at layer $l+1$. Dilated convolutions are extensions of these convolutions, that add spacing between the elements of the kernel k so that neighboring voxels at larger intervals are considered when computing the value for a voxel x in F_{l+1}. The level of dilation determines the stride between kernel elements (Fig. 8.2). CNNs with dilated convolutions have several advantages over CNNs with non-dilated, i.e. standard, convolutions. First, by stacking convolution layers with increasing levels of dilation, the receptive field for every voxel can be substantially extended at the cost of only few additional trainable parameters. Second, dilated convolution operations are translationally equivalent: the same multi-scale feature aggregation pyramid is applied at each location in the image. Hence, translating the image results in a translated version of the original output. Third, no downsampling layers are required to obtain large receptive fields and hence, high resolution label maps can be directly predicted by the network.

[1] http://segchd.csail.mit.edu

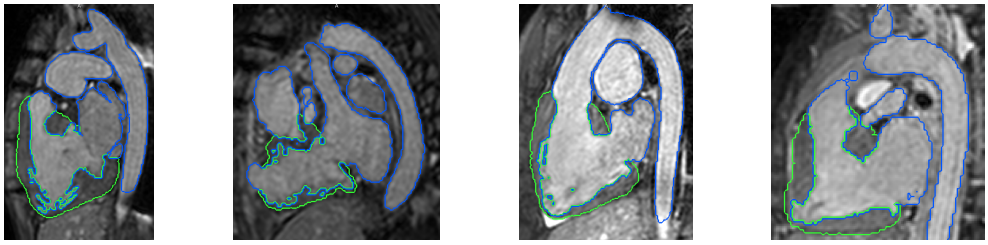


Figure 8.1: Example cardiovascular MR data of four patients with congenital heart disease. The examples illustrate the high variability in the structure and appearance of the blood pool and myocardium. Reference annotations for the blood pool and myocardium are shown in blue and green, respectively.

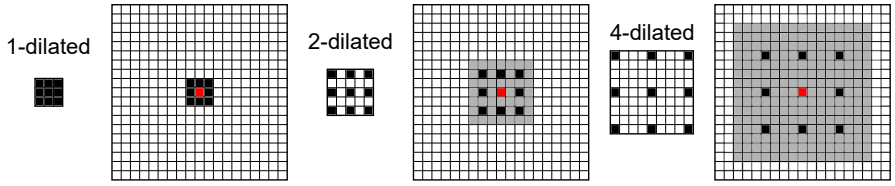


Figure 8.2: Convolution with a standard 1-dilated 3×3 kernel, followed by convolution with a 2-dilated and a 4-dilated kernel (kernels shown in black). The receptive field (shown in gray) increases from 3×3 after the first convolution, to 7×7 after the second convolution and 15×15 after the third convolution. All convolutions only use $3 \times 3 = 9$ trainable parameters.

Table 8.1: The convolutional neural network architecture used in this study. For each layer, the convolution kernel size, the level of dilation, the receptive field, the number of output channels and the number of trainable parameters are listed. chapter6Figures in the top row illustrate the receptive field at each layer shown in red.

										
Layer	1	2	3	4	5	6	7	8	9	10
Convolution	3×3	3×3	3×3	3×3	3×3	3×3	3×3	3×3	1×1	1×1
Dilation	1	1	2	4	8	16	32	1	1	1
Field	3×3	5×5	9×9	17×17	33×33	65×65	129×129	131×131	131×131	131×131
Channels	32	32	32	32	32	32	32	32	192	3
Parameters	320	9248	9248	9248	9248	9248	9248	9344	6912	579

Table 8.1 lists the CNN architecture used in this study. We adapt the dilated convolution context module proposed by Yu et al. [136] and extend the receptive field from 67×67 voxels to 131×131 voxels by inserting layer 7, with 32-dilated kernels. Layers 1 to 8 serve as feature extraction layers, while layers 9 and 10 are fully connected classification layers, implemented as 1×1 convolutional layers for increased efficiency. In each feature extraction layer, 32 kernels are used. The dila-

tion level is increased between layers 2 and 7. This allows the receptive field to grow exponentially, while the number of trainable parameters grows linearly; the same number of parameters is used in each layer. The Figures in the top row illustrate the receptive field in each layer. Layers 1 to 9 are each followed by an exponential linear unit (ELU) activation function [180], while layer 10 is followed by a softmax function. Batch normalization and dropout are applied to the fully connected layers 8 and 9 [137, 113]. Classification is performed in a wider layer with 192 channels and a final layer with 3 output channels, i.e. for the myocardium, blood pool and background. In total, the network contains 72,643 trainable parameters.

To correct for differences in intensity signals between CMR images, each image was normalized to have zero mean and unit variance. Furthermore, to correct for potential differences in orientation of patients, copies of the images rotated 90, 180 and 270 degrees along each axis were added to the training set. Finally, to guarantee that structures appear at similar scales in different images and along different axes, all images were resampled to an isotropic resolution of 0.65×0.65×0.65 mm per voxel, the smallest voxel dimension present in the data set. These isotropic volumes were used for voxel classification. A single CNN was trained to segment axial, sagittal and coronal image slices. During testing, full slices with 65-voxel zero-padding were processed so that for each viewing direction, 3D probabilistic maps were obtained for the myocardium, blood pool and background classes. These maps were averaged per segmentation class and resampled to the original input dimensions using trilinear interpolation. Finally, each voxel was assigned the segmentation class label with the highest posterior probability. Hence, the final probability for the myocardium, blood pool or background for each voxel depends on three 131×131 patches centered at that voxel. To guarantee contiguous myocardium and blood pool segmentations, only the largest component for each class was included in the final segmentation.

Evaluation was performed through an online system that was provided by the HVSMR challenge. The overlap between reference and automatically obtained segmentations was computed using the Dice index. Furthermore, the difference between reference and automatically obtained boundaries was computed using the average distance to boundaries (ADB) and the Hausdorff distance.

8.4 Experiments and results

We performed a five-fold cross-validation experiment on the training set, where each fold contained two CMR scans. Furthermore, to segment the test set, we trained a single CNN using all training images. Network parameters were optimized with Adam [181] using categorical cross-entropy as the cost function. Each CNN was trained with 10,000 training steps, which required 12 hours using a state-of-the-art GPU. In each training step a mini-batch containing 128 randomly selected 201×201 subimages from the training set was provided. Hence, in each training step the

Table 8.2: Results for the training and test set as provided by the HVSMR challenge. For both the myocardium and the blood pool, the Dice index, the average distance to boundaries (ADB) and the Hausdorff distance (Hausdorff) are listed.

		Myocardium			Blood pool		
		Dice	ADB	Hausdorff	Dice	ADB	Hausdorff
Training	Average	0.80 ± 0.06	1.01 ± 0.43	6.70 ± 3.52	0.92 ± 0.03	0.81 ± 0.28	5.86 ± 3.36
Test	Patient 10	0.72	1.34	10.75	0.94	0.74	5.23
	Patient 11	0.81	0.68	2.50	0.93	0.94	9.17
	Patient 12	0.87	0.60	3.94	0.93	0.83	9.74
	Patient 13	0.88	1.03	10.19	0.94	0.94	10.62
	Patient 14	0.71	1.33	8.69	0.90	1.07	4.21
	Patient 15	0.76	1.07	3.97	0.89	1.44	11.78
	Patient 16	0.76	0.80	3.14	0.91	0.77	6.12
	Patient 17	0.87	0.70	4.14	0.95	0.61	4.27
	Patient 18	0.85	0.61	2.19	0.94	0.64	3.29
	Patient 19	0.79	1.41	11.76	0.93	0.87	6.28
	Average	0.80 ± 0.06	0.96 ± 0.32	6.13 ± 3.76	0.93 ± 0.02	0.89 ± 0.24	7.07 ± 3.01

network optimized parameters for $71 \times 71 \times 128 = 645,248$ training voxels.

Table 8.2 lists the Dice index, the average distance to boundary (ADB), and the Hausdorff distance for automatic myocardium and blood pool segmentation in each of the ten test scans provided by the HVSMR challenge, as well as average scores for the training and test sets. Automatic segmentation of the *training* scans resulted in Dice indices of 0.80 ± 0.06 and 0.92 ± 0.03, average distances to boundaries of 1.01 ± 0.43 and 0.81 ± 0.28 mm, and Hausdorff distances of 6.70 ± 3.52 and 5.86 ± 3.36 mm for the myocardium and blood pool, respectively. Automatic segmentation of the *test* scans resulted in Dice indices of 0.80 ± 0.06 and 0.93 ± 0.02, average distances to boundaries of 0.96 ± 0.31 and 0.89 ± 0.24 mm, and Hausdorff distances of 6.13 ± 3.76 and 7.07 ± 3.01 mm for the myocardium and blood pool, respectively. The Dice index for myocardium was in all cases lower than the Dice index for blood pool segmentation. In several cases, Dice indices for the blood pool were affected by the (partial) identification of vessels that were not included in the reference standard, as shown in Fig. 8.3c. Segmentation of a CMR image took between 12.9 and 64.0 seconds, depending on image size, with an average of 41.5 ± 14.7 seconds.

To compare the performance of a CNN with dilated convolutions and a CNN without dilated convolutions, segmentations were performed using an otherwise identical CNN architecture containing 72,643 trainable parameters. Fig. 8.3 shows the obtained results. By omitting dilation, the receptive field for each voxel was reduced. While local information was used in classification by the CNN without dilation, long-range information was not, and hence the network was much less specific than the network with dilation.

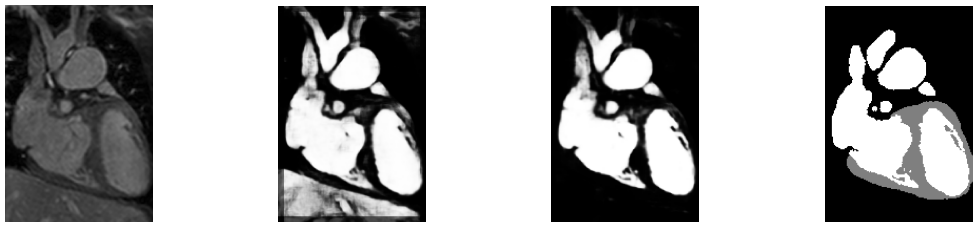


Figure 8.3: (a) Example CMR image slice. Probabilistic blood pool maps obtained using (b) a CNN without dilation (17×17 voxel receptive field) showing oversegmentation in the liver and (c) a CNN with dilation (131×131 voxel receptive field) showing no response in the liver. (d) reference annotation with the blood pool shown in white. Both networks have 72,643 trainable parameters.

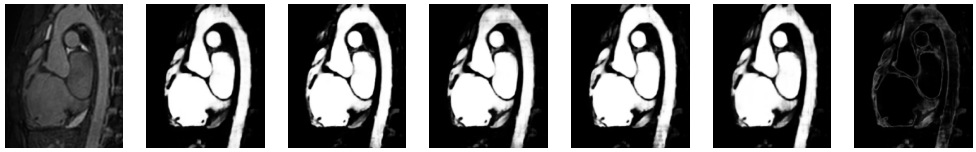


Figure 8.4: (a) CMR image slice, Patient 18. (b)-(f) Probabilistic blood pool maps obtained using CNNs trained on five different folds during cross-validation. (g) Standard deviation of probabilities predicted by CNNs.

To investigate whether overfitting may have ocurred in the five CNNs trained during cross-validation, we compared predictions made by these CNNs on an unseen image from the test set. Fig. 8.4 shows this image and probabilistic blood pool maps obtained by the CNNs. Even though each CNN was trained with only eight training images, the variance among the five predictions (Fig. 8.4g) was generally low, with higher values at the edges of the blood pool. Hence, it is unlikely that overfitting to the training data in each fold occurred.

8.5 Discussion and conclusion

We have presented a method for automatic segmentation of cardiovascular MR images in congenital heart disease using dilated convolutional neural networks. The method was able to accurately segment the myocardium and the blood pool without any expert intervention.

The current study showed that dilated convolution layers allow the combination of local structure and global context information with very few trainable parameters. Visual inspection of feature maps suggested that shallow layers enhanced

local image features such as edges, and deeper layers distinguished between locally similar but globally different areas. Our network used only 76,423 parameters, while comparable networks for medical image segmentation typically use more than 500,000 parameters [178]. This substantially reduces the risk of overfitting on the training data, which is particularly likely when training with very few scans as done in this study. In future work, we will investigate if the number of parameters can be further reduced without affecting performance, e.g. by reducing the number of output channels in each layer. The CNN was applied to full image slices to produce high-resolution output images, without downsampling of input or internal representations. We found that the method on average required only 41.5 seconds per scan, compared to 12.58 minutes in a recently published method for whole heart segmentation in cardiac MRI [182].

The method occasionally identified structures that were not included in the reference standard, but that are part of the blood pool, such as the distal sections of the descending aorta. It is unlikely that this will be problematic for the clinical purpose of segmentation of CMR in CHD patients. Dice indices for automatically obtained segmentations of the blood pool were in all patients higher than those of the myocardium. This is likely due to the lower image contrast between the myocardium and surrounding tissue, the more irregular shape of the myocardium and the difference in size between the myocardium and blood pool.

For each voxel, the final label depended on three orthogonal patches centered at that voxel. The information in these patches was combined in a late stage, by averaging of the three probabilities provided by the CNN. In future work, the features extracted from the three orthogonal patches may be fused before classification. In addition, we will investigate dilated convolutions in 3D, which might allow us to fully leverage the volumetric information present in the image. However, hardware limitations currently force a trade-off between dimensionality and receptive field size, i.e. it would be infeasible to train a 3D dilated CNN with $131 \times 131 \times 131$ receptive fields. Therefore, we have here chosen to use a larger receptive field at the cost of reduced volumetric information.

Chapter 9

Deep learning for multi-task medical image segmentation in multiple modalities

Based on:
P. Moeskops, J.M. Wolterink, B.H.M. van der Velden, K.G.A. Gilhuijs, T. Leiner, M.A. Viergever and I. Išgum "Deep learning for multi-task medical image segmentation in multiple modalities," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016, Lecture Notes in Computer Science, vol. 9901, pp. 478–486

Abstract

Automatic segmentation of medical images is an important task for many clinical applications. In practice, a wide range of anatomical structures are visualised using different imaging modalities. In this chapter, we investigate whether a single convolutional neural network (CNN) can be trained to perform different segmentation tasks.

A single CNN is trained to segment six tissues in MR brain images, the pectoral muscle in MR breast images, and the coronary arteries in cardiac CTA. The CNN therefore learns to identify the imaging modality, the visualised anatomical structures, and the tissue classes.

For each of the three tasks (brain MRI, breast MRI and cardiac CTA), this combined training procedure resulted in a segmentation performance equivalent to that of a CNN trained specifically for that task, demonstrating the high capacity of CNN architectures. Hence, a single system could be used in clinical practice to automatically perform diverse segmentation tasks without task-specific training.

9.1 Introduction

Automatic segmentation is an important task in medical images acquired with different modalities visualising a wide range of anatomical structures. A common approach to automatic segmentation is the use of supervised voxel classification, where a classifier is trained to assign a class label to each voxel. The classical approach to supervised classification is to train a classifier that discriminates between tissue classes based on a set of hand-crafted features. In contrast to this approach, convolutional neural networks (CNNs) automatically extract features that are optimised for the classification task at hand. CNNs have been successfully applied to medical image segmentation of e.g. knee cartilage [104], brain regions [183, 27], the pancreas [28], and coronary artery calcifications [62]. Each of these studies employed CNNs, but problem-specific optimisations with respect to the network architecture were still performed and networks were only trained to perform one specific task.

CNNs have not only been used for processing of medical images, but also for natural images. CNN architectures designed for image classification in natural images [90] have shown great generalisability for divergent tasks such as image segmentation [184], object detection [185], and object localisation in medical image analysis [103]. Hence, CNN architectures may have the flexibility to be used for different tasks with limited modifications.

In this study, we first investigate the feasibility of using a single CNN *architecture* for different medical image segmentation tasks in different imaging modalities visualising different anatomical structures. Secondly, we investigate the feasibility of using a single *trained instance* of this CNN architecture for different segmentation tasks. Such a system would be able to perform multiple tasks in different modalities without problem-specific tailoring of the network architecture or hyperparameters. Hence, the network recognises the modality of the image, the anatomy visualised in the image, and the tissues of interest. We demonstrate this concept using three different and potentially adversarial medical image segmentation problems: segmentation of six brain tissues in brain MRI, pectoral muscle segmentation in breast MRI, and coronary artery segmentation in cardiac CT angiography (CTA).

9.2 Data

Brain MRI – 34 T_1-weighted MR brain images from the OASIS project [186] were acquired on a Siemens Vision 1.5 T scanner, as provided by the MICCAI challenge on multi-atlas labelling [187][1]. The images were acquired with voxel sizes of $1.0 \times 1.0 \times 1.25$ mm^3 and resampled to isotropic voxel sizes of $1.0 \times 1.0 \times 1.0$ mm^3. The images were manually segmented, in the coronal plane, into 134 classes that were, for the purpose of this chapter, combined into six commonly used tissue classes: white mat-

[1] https://masi.vuse.vanderbilt.edu/workshop2012

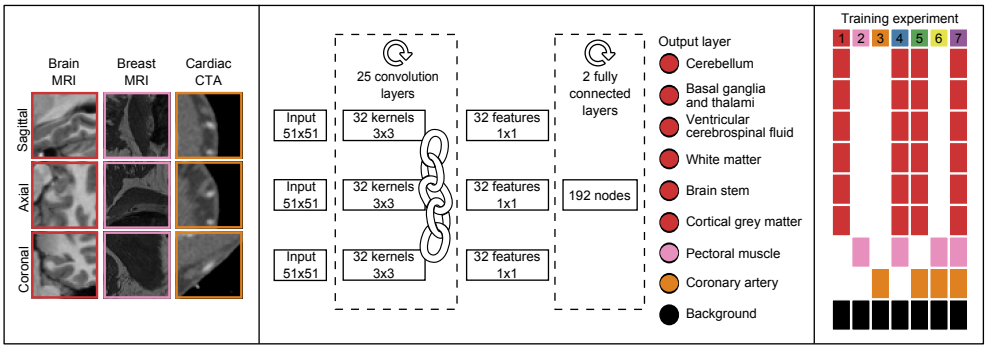


Figure 9.1: Example 51×51 triplanar input patches (*left*). CNN architecture with 25 shared convolution layers, 2 fully connected layers and an output layer with at most 9 classes, including a background class common among tasks (*centre*). Output classes included in each training experiment (*right*).

ter, cortical grey matter, basal ganglia and thalami, ventricular cerebrospinal fluid, cerebellum, and brain stem.

Breast MRI – 34 T_1-weighted MR breast images were acquired on a Siemens Magnetom 1.5 T scanner with a dedicated double breast array coil [188]. The images were acquired with in-plane voxel sizes between 1.21 and 1.35 mm and slice thicknesses between 1.35 and 1.69 mm. All images were resampled to isotropic voxel sizes corresponding to their in-plane voxel size. The pectoral muscle was manually segmented in the axial plane by contour drawing.

Cardiac CTA – Ten cardiac CTA scans were acquired on a 256-detector row Philips Brilliance iCT scanner using 120 kVp and 200-300 mAs, with ECG-triggering and contrast enhancement. The reconstructed images had between 0.4 and 0.5 mm in-plane voxel sizes and 0.45/0.90 mm slice spacing/thickness. All images were resampled to isotropic 0.45×0.45×0.45 mm^3 voxel size. To set a manual reference standard, a human observer traversed the scan in the craniocaudal direction and painted voxels in the main coronary arteries and their branches in the axial plane.

9.3 Method

All voxels in the images were labelled by a CNN using seven different training experiments (Fig. 9.1).

9.3.1 CNN architecture

For each voxel, three orthogonal (axial, sagittal, and coronal) patches of 51×51 voxels centred at the target voxel were extracted. For each of these three patches, fea-

tures were determined using a deep stack of convolution layers. Each convolution layer contained 32 small (3×3 voxels) convolution kernels for a total of 25 convolution layers [108]. To prevent over- or undersegmentation of structures due to translational invariance, no subsampling layers were used. To reduce the number of trainable parameters in the network and hence the risk of over-fitting, the same stack of convolutional layers was used for the axial, sagittal and coronal patches.

The output of the convolution layers were 32 features for each of the three orthogonal input patches, hence, 96 features in total. These features were input to two subsequent fully connected layers, each with 192 nodes. The second fully connected layer was connected to a softmax classification layer. Depending on the tasks of the network, this layer contained 2, 7, 8 or 9 output nodes. The fully connected layers were implemented as 1×1 voxel convolutions, to allow fast processing of arbitrarily sized images. Exponential linear units [180] were used for all non-linear activation functions. Batch normalisation [137] was used on all layers and dropout [113] was used on the fully connected layers.

9.3.2 Training experiments

The same model was trained for each combination of the three tasks. In total seven training experiments were performed (Fig. 9.1, right): three networks were trained to perform one task (Experiments 1–3), three networks were trained to perform two tasks (Experiments 4–6), and one network was trained to perform three tasks (Experiment 7). The number of output nodes in the CNN was modified accordingly. In each experiment, background classes of the target tasks were merged into one class.

Each CNN was trained using mini-batch learning. A mini-batch contained 210 samples, equally balanced over the tasks of the network. For each task, the training samples were randomly drawn from all training images, balanced over the task-specific classes. All voxels with image intensity > 0 were considered samples. The network parameters were optimized using Adam stochastic optimisation [181] with categorical cross-entropy as the cost-function.

9.4 Experiments and results

The data for brain MRI, breast MRI and cardiac CTA were split into 14/20, 14/20 and 6/4 training/test images, respectively. Four results were obtained for each task: one with a network trained for only that task, two with networks trained for that task and an additional task, and one with a network trained for all tasks together. Each network was trained with 25000 mini-batches per task.

No post-processing steps other than probability thresholding for evaluation purposes were performed. The results are presented on the full test set. In brain MRI, the voxel class labels were determined by the highest class activation. The performance was evaluated per brain tissue type, using the Dice coefficient between the

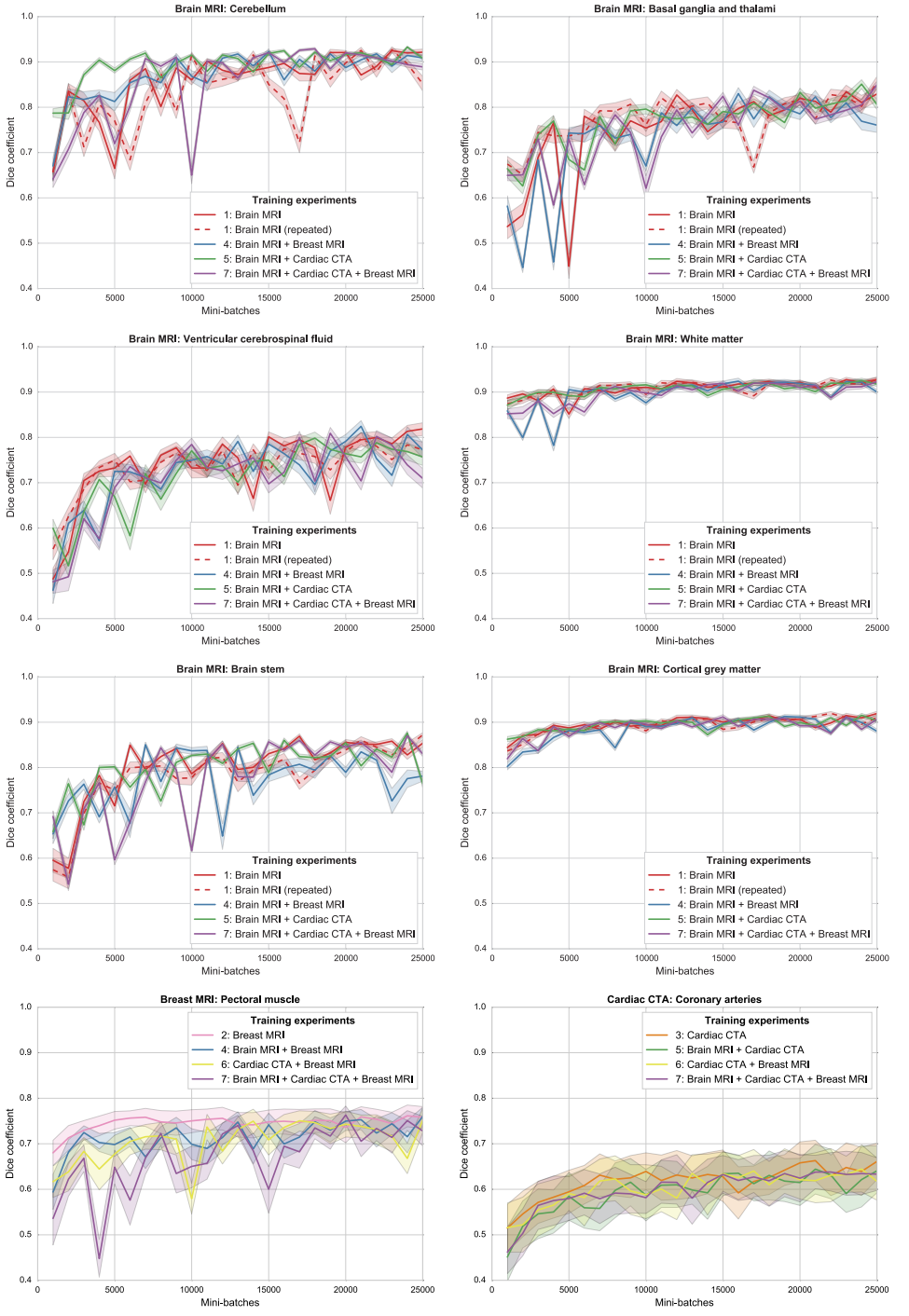


Figure 9.2: Learning curves showing Dice coefficients for tissue segmentation in brain MRI (*top three rows*), breast MRI (*bottom left*), and cardiac CTA (*bottom right*), reported at 1000 mini-batch intervals for experiments including that task. The line colours correspond to the training experiments in Fig. 9.1.

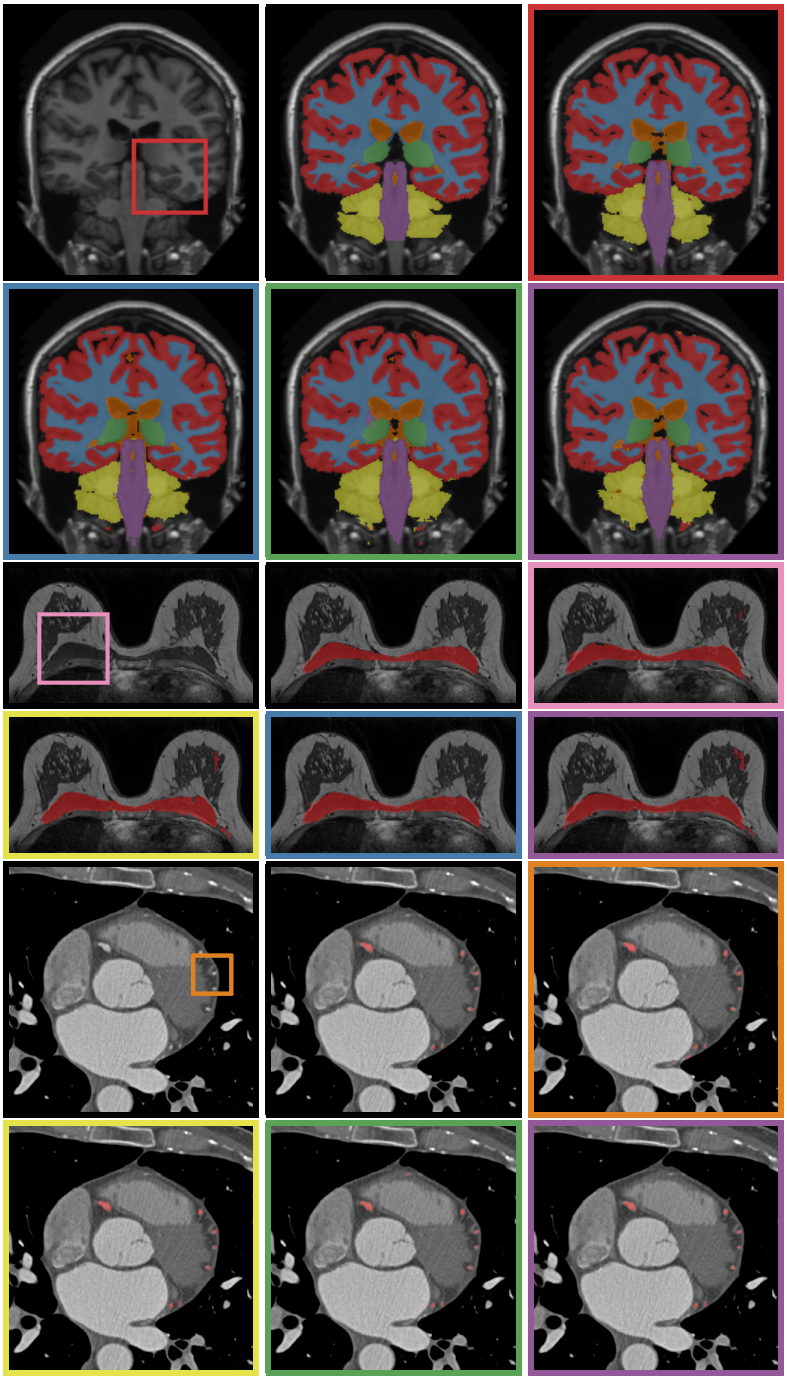


Figure 9.3: Example segmentations for (*top to bottom*) brain MRI, breast MRI, and cardiac CTA. Shown for each task: (*left to right*, *first row*) image with an input patch as shown in Fig. 9.1, reference standard, segmentation by task-specific training, (*left to right*, *second row*) two segmentations by networks with an additional task, segmentation by a network combining all tasks. The coloured borders correspond to the training experiments in Fig. 9.1 and Fig. 9.2.

manual and automatic segmentations. In breast MRI and cardiac CTA, precision-recall curve analysis was performed to identify the optimal operating point, defined, for each experiment, as the highest Dice coefficient over the whole test set. The thresholds at this optimal operating point were then applied to all images.

Fig. 9.2 shows the results of the described quantitative analysis, performed at intervals of 1000 mini-batches per task. As the networks learned, the obtained Dice coefficients increased and the stability of the results improved. For each segmentation task, the learning curves were similar for all experiments. Nevertheless, slight differences were visible between the obtained learning curves. To assess whether these differences were systematic or caused by the stochastic nature of CNN training, the training experiment using only brain MR data (Experiment 1) was repeated (dashed line in Fig. 9.2), showing similar inter-experiment variation. Fig. 9.3 shows a visual comparison of results obtained for the three different tasks. For all three tasks, all four networks were able to accurately segment the target tissues.

Confusion between tasks was very low. For the network trained with three tasks, the median percentage of voxels per scan labelled with a class alien to the target (e.g. cortical grey matter identified in breast MR) was $< 0.0005\%$ for all tasks.

9.5 Discussion and conclusions

The results demonstrate that a single CNN architecture can be used to train CNNs able to obtain accurate results in images from different modalities, visualising different anatomy. Moreover, it is possible to train a single CNN instance that can not only segment multiple tissue classes in a single modality visualising a single anatomical structure, but also multiple classes over multiple modalities visualising multiple anatomical structures.

In all experiments, a fixed CNN architecture with triplanar orthogonal input patches was used. We have strived to utilise recent advances in deep learning such as batch normalisation [137], Adam stochastic optimisation [181], exponential linear units [180], and very deep networks with small convolution kernels [108]. Furthermore, the implementation of fully connected layers as 1×1 convolution layers and the omission of downsampling layers allowed fast processing of whole images compared with more time-consuming patch-based scanning [104, 28, 62, 27]. The ability of the CNN to adapt to different tasks suggests that small architectural changes are unlikely to have a large effect on the performance. Volumetric 3D input patches might result in increased performance, but would require a high computational load due to the increased size of the network parameter space.

The results for brain segmentation are comparable with previously published results [27]. Due to differences in image acquisition and patient population, the obtained results for pectoral muscle segmentation and coronary artery extraction cannot be directly compared to results reported in other studies. Nevertheless, these results appear to be in line with previously published studies [189, 190]. No post-

processing other than probability thresholding for evaluation purposes was applied. The output probabilities may be further processed, or directly used as input for further analysis, depending on the application.

Including multiple tasks in the training procedure resulted in a segmentation performance equivalent to that of a network trained specifically for the task (Fig. 9.2). Similarities between the tasks, e.g. presence of the pectoral muscle in both breast MR and cardiac CTA, or similar appearance of brain and breast tissue in T_1-weighted MRI, led to very limited confusion. In future work, we will further investigate the capacity of the current architecture with more data and segmentation tasks, and investigate to what extent the representations within the CNN are shared between tasks.

Chapter 10

Summary and discussion

This chapter provides a summary and discussion of the work presented in this thesis.

10.1 Summary

Chapter 2 presented a method for automatic coronary calcium scoring in non-contrast-enhanced cardiac CT scans. Candidate calcifications are described using a set of shape, texture and location features. A forest of extremely randomized trees identifies coronary calcifications and labels these according to the artery they are in. Uncertainty of the classifier is quantified using the entropy in its posterior probability distribution. Experiments showed that half of the analyzed scans contained candidates with an entropy indicating uncertainty. Presentation of only these candidates to an expert substantially improved performance. This allows both accurate and fast coronary calcium scoring.

Chapter 3 described an evaluation framework for (semi-)automatic coronary calcium scoring methods in non-contrast-enhanced cardiac CT scans. The framework provides CT scans of patients assigned to different CVD risk categories acquired on four different scanners from four different vendors. A standardized evaluation of five (semi-)automatic methods within this framework showed that automatic per patient CVD risk categorization is feasible. CAC lesions at ambiguous locations such as the coronary ostia remain challenging, but their detection had limited impact on the accuracy of CVD risk determination.

Chapter 4 described an efficient method for 3D organ localization in medical image volumes. A convolutional neural network is trained to identify presence of an organ of interest in 2D image slices. Responses along the three axes of a 3D volume are combined into a bounding box around the organ. The method was able to identify the heart, aortic arch and descending aorta in non-contrast-enhanced chest CT, with excellent results and short processing times. This is a useful preprocessing application for many medical image analysis algorithms.

Chapter 5 presented a deep learning-based method for automatic coronary calcium scoring in contrast-enhanced cardiac CT. The image is first cropped around the heart using the method described in Chapter 4 and a convolutional neural network is used to identify candidate CAC voxels in the cropped volume. A second network with an identical architecture is used to identify real CAC voxels among these candidates. The results showed that this cascaded scheme facilitates classification with extreme class imbalance. Automatically determined calcium mass scores showed a strong correlation with manual calcium mass scores in non-contrast-enhanced CT. Hence, automatic calcium scoring in contrast-enhanced CT might obviate the need for a dedicated non-contrast CT scan for CAC scoring, and thus reduce the CT radiation dose received by patients.

Chapter 6 described a method for coronary centerline extraction from contrast-enhanced cardiac CT scans. In contrast to previously published methods, no pre-

defined vesselness filter is used. Instead, a convolutional neural network is trained to simultaneously predict the direction and radius of the coronary artery, based on only CT values. The accuracy of these predictions was demonstrated in an iterative tracking algorithm, as well as in a comparison of reference and automatically determined artery radius values. This method only requires annotated centerlines for training, and could be adapted to other vessels and modalities.

Chapter 7 described a method for noise reduction in low-dose CT. A convolutional neural network is trained to transform a low-dose CT image into a reference routine-dose CT image. To prevent the network from generating smooth images, feedback from an adversarial discriminator network is used. The results showed that the images produced by networks with adversarial training are more similar to the routine-dose image than those produced by networks without adversarial training. Furthermore, noise reduction allowed calcium quantification in low-dose patient cardiac CT images in which calcium scoring was otherwise infeasible.

Chapter 8 presented a convolutional neural network for segmentation of the myocardium and blood pool in cardiovascular MR images of congenital heart disease patients. The network uses dilated convolution kernels, which provide high-resolution feature maps, and reduce the number of parameters in the network. The method achieved top-ranking results among evaluated automatic methods in the MICCAI 2016 Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease (HVSMR).

Chapter 9 described a series of experiments in which a single convolutional neural network was trained to perform coronary artery segmentation in cardiac CT, brain tissue segmentation in brain MRI and pectoral muscle segmentation in breast MRI. The results indicated that the network identified the imaging modality and anatomy being visualized. Performance on any of the three tasks was not hampered by their combination, and there was very little confusion between tasks. This demonstrates the large capacity of convolutional neural networks, and may be a stepping stone to a more general system in which more segmentation tasks and imaging modalities are combined.

10.2 Discussion and future perspectives

Chapters 2, 3 and 5 of this thesis showed that coronary calcium can be quantified in non-contrast-enhanced cardiac CT, as well as in contrast-enhanced CT scans. In clinical practice both scans are typically acquired, but the non-contrast-enhanced CT scan is used only for calcium scoring. Therefore, this acquisition could potentially be omitted to reduce the ionizing radiation dose received by patients. However, manual calcium scoring in contrast-enhanced CT is considerably more time-consuming. Contrast-enhanced CT scans typically contain up to 300 slices, compared to around 50 slices in non-contrast CT. Furthermore, while there have been large-scale studies on the relation between CAC in non-contrast CT and cardiovascular risk [10, 191],

such studies have not been performed for coronary calcium in contrast-enhanced CT. Therefore, contrast-enhanced CT calcium scores are typically first converted to non-contrast scores for risk prediction [192]. Automatic methods like the one proposed in Chapter 5 could alleviate these problems, and allow large-scale studies into cardiovascular risk prediction from contrast-enhanced CT.

Besides calcium scoring, contrast-enhanced CT can be used for tasks requiring segmentation of the coronary artery lumen, like automatic stenosis detection [193] and computation of CT-based fractional flow reserve measurements [194]. Chapter 6 described a method to extract the centerline of the coronary lumen along with an estimate of its radius. This method only processes patches along the centerline and thus has a small computational footprint. A complementary method proposed in Chapter 9 uses a convolutional neural network to provide voxel-wise segmentations of the coronary arteries, based on segmentation of the full image. In future work, these methods could be combined to rapidly provide accurate coronary lumen segmentations.

All methods discussed in this thesis are supervised, and are prone to overfitting when insufficient labeled training samples are available. This is particularly likely in large deep learning models with millions of parameters. Therefore, the method described in Chapter 8 aimed to reduce the number of parameters with dilated convolution layers [136]. Such convolution layers provide large receptive fields and high-resolution predictions with few layers and parameters. Alternatively, overfitting may be prevented by data augmentation as in Chapter 6, or with techniques like dropout [113] and batch normalization [137], which were routinely used in this work.

Most methods in medical image analysis focus on a single task in a single imaging modality. However, it has been shown that representations learned by neural networks can be shared between tasks, even when these tasks are in seemingly different domains [29, 30, 31]. In the clinical workflow, a system that automatically provides a segmentation of every newly acquired image, irrespective of its modality or the anatomy visualized, would be highly useful. In Chapter 9, a stepping stone towards such a system was presented. The results showed that one CNN could be trained to segment cardiac CT, breast MRI and brain MRI. However, unlike results reported for multi-task learning in e.g. robotics [195], the performance on single tasks did not appear to be improved by their combination with other tasks. In future work, neural network representation sharing in multi-task learning will be further investigated. In addition, the system may be adapted to sequentially learn new tasks. To this end, techniques to prevent catastrophic forgetting of previously learned tasks could be integrated [196].

Automatic machine learning methods are expected to have a large influence on medical fields which include repetitive and scalable tasks, such as histopathology, dermatology and radiology[197, 198]. Commonly cited advantages of automated methods are their consistent performance and low costs. For example, near human-

level performance has recently been reported on diabetic retinopathy [130] and dermatological detection [31]. However, many tasks can still benefit from human expertise. An analysis of automatic calcium scoring methods in Chapter 3 showed that classification of candidate calcifications near the coronary ostia remains challenging. To this end, the method described in Chapter 2 allows efficient collaboration with an expert. While fully automatic CAC scoring assigned 93% of patients to their reference CVD risk category, guided review by an expert improved this number to 99% with minimal effort. In such applications, it is important that the automatic method can select the right samples for review, such as large calcified lymph nodes in calcium scoring (Chapter 5) or an usually large aorta in organ localization (Chapter 4). Recently proposed methods [199, 200] may help better quantify the uncertainty of machine learning methods. Expert decisions on reviewed samples could be fed back to the machine learning system so that it continuously improves.

The methods in this thesis focused on coronary CT and cardiac MR. However, these techniques may also be applied to other cardiovascular imaging modalities. For example, the organ detection technique in Chapter 4 could be used in any 3D medical imaging volume, and the centerline extraction method in Chapter 6 may be applied to other vasculatures in cardiovascular images. Furthermore, recent developments indicate that MR imaging of the coronary arteries and atherosclerotic plaque may be feasible in the near future [201]. In future work, it would be interesting to investigate how the techniques described in this thesis translate to coronary MR imaging and other cardiovascular imaging modalities.

Bibliography

[1] World Health Organization, "WHO Fact sheet No. 317: Cardiovascular Diseases," 2013.

[2] G. K. Hansson, "Inflammation, atherosclerosis, and coronary artery disease," *N. Engl. J. Med.*, vol. 352, no. 16, pp. 1685–1695, 2005.

[3] J. F. Bentzon, F. Otsuka, R. Virmani, and E. Falk, "Mechanisms of plaque formation and rupture," *Circ. Res.*, vol. 114, no. 12, pp. 1852–1866, 2014.

[4] R. J. Myerburg and M. J. Junttila, "Sudden cardiac death caused by coronary heart disease," *Circulation*, vol. 125, no. 8, pp. 1043–1052, 2012.

[5] A. Mauriello, F. Servadei, G. B. Zoccai, E. Giacobbi, L. Anemona, E. Bonanno, and S. Casella, "Coronary calcification identifies the vulnerable patient rather than the vulnerable plaque," *Atherosclerosis*, vol. 229, no. 1, pp. 124–129, 2013.

[6] F. Otsuka, K. Sakakura, K. Yahagi, M. Joner, and R. Virmani, "Has our understanding of calcification in human coronary atherosclerosis progressed?," *Arterioscler. Thromb. Vasc. Biol.*, vol. 34, no. 4, pp. 724–736, 2014.

[7] R. Detrano, A. D. Guerci, J. J. Carr, D. E. Bild, G. Burke, A. R. Folsom, K. Liu, S. Shea, M. Szklo, D. A. Bluemke, *et al.*, "Coronary calcium as a predictor of coronary events in four racial or ethnic groups," *N. Engl. J. Med.*, vol. 358, no. 13, pp. 1336–1345, 2008.

[8] J. Yeboah, R. L. McClelland, T. S. Polonsky, G. L. Burke, C. T. Sibley, D. O'Leary, J. J. Carr, D. C. Goff, P. Greenland, and D. M. Herrington, "Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals," *JAMA*, vol. 308, no. 8, pp. 788–795, 2012.

[9] 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," *J. Am. Coll. Cardiol.*, vol. 15, no. 4, pp. 827–832, 1990.

[10] J. A. Rumberger, B. H. Brundage, D. J. Rader, and G. Kondos, "Electron beam computed tomographic coronary calcium scanning: A review and guidelines for use in asymptomatic persons," *Mayo Clin. Proc.*, vol. 74, no. 3, pp. 243–252, 1999.

[11] C. H. McCollough, S. Ulzheimer, S. S. Halliburton, K. Shanneik, R. D. White, and W. A. Kalender, "Coronary artery calcium: A multi-institutional, multimanufacturer international standard for quantification at cardiac CT," *Radiology*, vol. 243, no. 2, pp. 527–538, 2007.

[12] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. O'Donnell, J. G. Robinson, J. S. Schwartz, S. T. Shero, S. C. Smith, P. Sorlie, N. J. Stone, and P. W. F. Wilson, "2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American College of Cardiology/American Heart Association task force on practice guidelines," *J. Am. Coll. Cardiol.*, vol. 63, no. 25, pp. 2935–2959, 2014.

[13] R. Shahzad, T. van Walsum, M. Schaap, A. Rossi, S. Klein, A. C. Weustink, P. J. de Feyter, L. J. van Vliet, and W. J. Niessen, "Vessel specific coronary artery calcium scoring: An automatic system," *Acad. Radiol.*, vol. 20, no. 1, pp. 1–9, 2013.

[14] I. Išgum, M. Prokop, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Automatic coronary calcium scoring in low-dose chest computed tomography," *IEEE Trans Med Imag*, vol. 31, no. 12, pp. 2322–2334, 2012.

[15] J. M. Wolterink, T. Leiner, R. A. P. Takx, M. A. Viergever, and I. Išgum, "Automatic coronary calcium scoring in non-contrast-enhanced ECG-triggered cardiac CT with ambiguity detection," *IEEE Trans Med Imag*, vol. 34, no. 9, pp. 1867–1878, 2015.

[16] N. Lessmann, I. Išgum, A. A. Setio, B. D. de Vos, F. Ciompi, P. A. de Jong, M. Oudkerk, P. T. M. Willem, M. A. Viergever, and B. van Ginneken, "Deep convolutional neural networks for automatic coronary calcium scoring in a screening study with low-dose chest CT," in *Proc. SPIE Med. Imag.*, vol. 9785, p. 978511, 2016.

[17] D. J. Brenner and E. J. Hall, "Computed tomography - an increasing source of radiation exposure," *N. Engl. J. Med.*, vol. 357, no. 22, pp. 2277–2284, 2007.

[18] M. S. Pearce, J. A. Salotti, M. P. Little, K. McHugh, C. Lee, K. P. Kim, N. L. Howe, C. M. Ronckers, P. Rajaraman, A. W. Craft, *et al.*, "Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: A retrospective cohort study," *The Lancet*, vol. 380, no. 9840, pp. 499–505, 2012.

[19] C. W. Pavitt, K. Harron, A. C. Lindsay, R. Ray, S. Zielke, D. Gordon, M. B. Rubens, S. P. Padley, and E. D. Nicol, "Deriving coronary artery calcium scores from CT coronary angiography: A proposed algorithm for evaluating stable chest pain," *Int. J. Cardiovasc. Imaging*, vol. 30, no. 6, pp. 1135–1143, 2014.

[20] I. Mylonas, M. Alam, N. Amily, G. Small, L. Chen, Y. Yam, B. Hibbert, and B. J. Chow, "Quantifying coronary artery calcification from a contrast-enhanced cardiac computed tomography angiography study," *Eur Heart J Cardiovasc Imaging*, vol. 15, no. 2, pp. 210–215, 2014.

[21] S. M. Gilboa, O. J. Devine, J. E. Kucik, M. E. Oster, T. Riehle-Colarusso, W. N. Nembhard, P. Xu, A. Correa, K. Jenkins, and A. J. Marelli, "Congenital heart defects in the United States: Estimating the magnitude of the affected population in 2010," *Circulation*, vol. 134, no. 2, pp. 101–109, 2016.

[22] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.

[23] I. Valverde, G. Gomez, A. Gonzalez, C. Suarez-Mejias, A. Adsuar, J. F. Coserria, S. Uribe, T. Gomez-Cia, and A. R. Hosseinpour, "Three-dimensional patient-specific cardiac model for surgical planning in Nikaidoh procedure," *Cardiol. Young*, vol. 25, no. 4, pp. 698–704, 2015.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] 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 Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV 2014, Part I* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), vol. 8689 of *LNCS*, pp. 818–833, Springer International Publishing, 2014.

[27] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Trans Med Imag*, vol. 35, no. 5, pp. 1252–1261, 2016.

[28] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *MICCAI 2015, Part I* (N. Navab, J. Hornegger, M. W. Wells, and F. A. Frangi, eds.), vol. 9349 of *LNCS*, pp. 556–564, Springer, 2015.

[29] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *CVPR Workshops 2014*, pp. 806–813, 2014.

[30] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, and B. van Ginneken, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box," *Med. Image Anal.*, vol. 26, no. 1, pp. 195–202, 2015.

[31] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

[32] T. Q. Callister, B. Cooil, S. P. Raya, N. J. Lippolis, D. J. Russo, and P. Raggi, "Coronary artery disease: Improved reproducibility of calcium scoring with an electron-beam CT volumetric method," *Radiology*, vol. 208, no. 3, pp. 807–814, 1998.

[33] I. Išgum, A. Rutten, M. Prokop, and B. van Ginneken, "Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease," *Med. Phys.*, vol. 34, no. 4, pp. 1450–1461, 2007.

[34] U. Kurkure, D. R. Chittajallu, G. Brunner, Y. H. Le, and I. A. Kakadiaris, "A supervised classification-based method for coronary calcium detection in non-contrast CT," *Int. J. Cardiovasc. Imaging*, vol. 26, no. 7, pp. 817–828, 2010.

[35] G. Brunner, D. R. Chittajallu, U. Kurkure, and I. A. Kakadiaris, "Toward the automatic detection of coronary artery calcification in non-contrast computed tomography data," *Int. J. Cardiovasc. Imaging*, vol. 26, no. 7, pp. 829–838, 2010.

[36] S. C. Saur, H. Alkadhi, L. Desbiolles, G. Székely, and P. C. Cattin, "Automatic detection of calcified coronary plaques in computed tomography data sets," in *MICCAI 2008, Part I* (D. Metaxas, L. Axel, G. Fichtinger, and G. Székely, eds.), vol. 5241 of *LNCS*, pp. 170–177, Springer Berlin Heidelberg, 2008.

[37] Y. Xie, M. D. Cham, C. Henschke, D. Yankelevitz, and A. P. Reeves, "Automated coronary artery calcification detection on low-dose chest CT images," in *Proc. SPIE Med. Imag.*, vol. 9035, p. 90350F, 2014.

[38] M. J. Budoff, J. E. Hokanson, K. Nasir, L. J. Shaw, G. L. Kinney, D. Chow, D. DeMoss, V. Nuguri, V. Nabavi, R. Ratakonda, *et al.*, "Progression of coronary artery calcium

predicts all-cause mortality," *JACC Cardiovasc Imaging*, vol. 3, no. 12, pp. 1229–1236, 2010.

[39] J. R. Ghadri, R. Goetti, M. Fiechter, A. P. Pazhenkottil, S. M. Küest, R. N. Nkoulou, C. Windler, R. R. Buechel, B. A. Herzog, O. Gaemperli, C. Templin, and P. A. Kaufmann, "Inter-scan variability of coronary artery calcium scoring assessed on 64-multidetector computed tomography vs. dual-source computed tomography: A head-to-head comparison," *Eur. Heart J.*, vol. 32, no. 15, pp. 1865–1874, 2011.

[40] J. M. Wolterink, T. Leiner, R. A. P. Takx, M. A. Viergever, and I. Išgum, "An automatic machine learning system for coronary calcium scoring in clinical non-contrast enhanced, ECG-triggered cardiac CT," in *Proc. SPIE Med. Imag.*, vol. 9035, p. 90350E, 2014.

[41] E. M. Urbina, R. V. Williams, B. S. Alpert, R. T. Collins, S. R. Daniels, L. Hayman, M. Jacobson, L. Mahoney, M. Mietus-Snyder, A. Rocchini, *et al.*, "Noninvasive assessment of subclinical atherosclerosis in children and adolescents: Recommendations for standard assessment for clinical research: A scientific statement from the American Heart Association," *Hypertension*, vol. 54, no. 5, pp. 919–950, 2009.

[42] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: A toolbox for intensity-based medical image registration," *IEEE Trans Med Imag*, vol. 29, no. 1, pp. 196–205, 2010.

[43] T. R. Langerak, U. A. van der Heide, A. N. Kotte, M. A. Viergever, M. van Vulpen, and J. P. W. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *IEEE Trans Med Imag*, vol. 29, no. 12, pp. 2000–2008, 2010.

[44] E. Weiszfeld and F. Plastria, "On the point for which the sum of the distances to n given points is minimum," *Ann Oper Res*, vol. 167, no. 1, pp. 7–41, 2009.

[45] A. Feragen, P. Lo, M. de Bruijne, M. Nielsen, and F. Lauze, "Toward a theory of statistical tree-shape analysis," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 8, pp. 2008–2021, 2013.

[46] T. Eiter and H. Mannila, "Computing discrete Frechet distance," Tech. Rep. CD-TR 94/64, Christian Doppler Laboratory for Expert Systems, 1994.

[47] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, no. 1, pp. 3–42, 2006.

[48] N. Lachiche and P. Flach, "Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves," in *ICML 2003*, pp. 416–423, 2003.

[49] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. SIGIR*, pp. 3–12, Springer-Verlag New York, Inc., 1994.

[50] A. Sarwar, L. J. Shaw, M. D. Shapiro, R. Blankstein, U. Hoffman, R. C. Cury, S. Abbara, T. J. Brady, M. J. Budoff, R. S. Blumenthal, *et al.*, "Diagnostic and prognostic value of absence of coronary artery calcification," *JACC Cardiovasc Imaging*, vol. 2, no. 6, pp. 675–688, 2009.

[51] M. J. Pletcher, M. Pignone, S. Earnshaw, C. McDade, K. A. Phillips, R. Auer, L. Zablotska, and P. Greenland, "Using the coronary artery calcium score to guide statin therapy: A cost-effectiveness analysis," *Circ Cardiovasc Qual Outcomes*, vol. 7, no. 2, pp. 276–284, 2014.

[52] M. S. Bittencourt, M. J. Blaha, R. Blankstein, M. Budoff, J. D. Vargas, R. S. Blumenthal, A. S. Agatston, and K. Nasir, "Polypill therapy, subclinical atherosclerosis, and cardiovascular events - Implications for the use of preventive pharmacotherapy: MESA (Multi-Ethnic Study of Atherosclerosis)," *J. Am. Coll. Cardiol.*, vol. 63, no. 5, pp. 434–443, 2014.

[53] W. Fan, E. Greengrass, J. McCloskey, P. S. Yu, and K. Drummey, "Effective estimation of posterior probabilities: Explaining the accuracy of randomized decision tree approaches," in *Int Conf Data Mining*, pp. 154–161, IEEE Comp. Soc., 2005.

[54] M. Schaap, C. T. Metz, T. van Walsum, A. G. van der Giessen, A. C. Weustink, N. R. Mollet, C. Bauer, H. Bogunović, C. Castro, X. Deng, *et al.*, "Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms," *Med. Image Anal.*, vol. 13, no. 5, pp. 701–714, 2009.

[55] P. Medrano-Gracia, J. Ormiston, M. Webster, S. Beier, C. Ellis, C. Wang, A. A. Young, and B. R. Cowan, "Construction of a coronary artery atlas from ct angiography," in *MICCAI 2014, Part II* (P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, eds.), vol. 8674 of *LNCS*, pp. 513–520, Springer International Publishing, 2014.

[56] M. J. Willemink, R. Vliegenthart, R. A. Takx, T. Leiner, R. P. Budde, R. L. Bleys, M. Das, J. E. Wildberger, M. Prokop, N. Buls, *et al.*, "Coronary artery calcification scoring with state-of-the-art CT scanners from different vendors has substantial effect on risk classification," *Radiology*, vol. 273, no. 3, pp. 695–702, 2014.

[57] A. Becker, A. Leber, C. Becker, and A. Knez, "Predictive value of coronary calcifications for future cardiac events in asymptomatic individuals," *A. Heart J.*, vol. 155, no. 1, pp. 154–160, 2008.

[58] K. Nasir and M. Clouse, "Role of nonenhanced multidetector CT coronary artery calcium testing in asymptomatic and symptomatic individuals," *Radiology*, vol. 264, no. 3, pp. 637–649, 2012.

[59] B. A. Arnold, P. Xiang, M. J. Budoff, and S. S. Mao, "Very small calcifications are detected and scored in the coronary arteries from small voxel MDCT images using a new automated/calibrated scoring method with statistical and patient specific plaque definitions," *Int. J. Cardiovasc. Imaging*, vol. 28, no. 5, pp. 1193–1204, 2012.

[60] X. Ding, P. J. Slomka, M. Diaz-Zamudio, G. Germano, D. S. Berman, D. Terzopoulos, and D. Dey, "Automated coronary artery calcium scoring from non-contrast CT using a patient-specific algorithm," in *Proc. SPIE Med. Imag.*, vol. 9413, p. 94132U, 2015.

[61] J. Wu, G. Ferns, J. Giles, and E. Lewis, "A fully automated multi-modal computer aided diagnosis approach to coronary calcium scoring of MSCT images," in *Proc. SPIE Med. Imag.*, vol. 8315, p. 83152I, 2012.

[62] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Automatic coronary calcium scoring in cardiac CT angiography using convolutional neural networks," in *MICCAI 2015, Part I* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), vol. 9349 of *LNCS*, pp. 589–596, Springer International Publishing, 2015.

[63] W. Ahmed, M. A. de Graaf, A. Broersen, P. H. Kitslaar, E. Oost, J. Dijkstra, J. J. Bax, J. H. Reiber, and A. J. Scholte, "Automatic detection and quantification of the Agatston coronary artery calcium score on contrast computed tomography angiography," *Int. J. Cardiovasc. Imaging*, vol. 31, no. 1, pp. 151–161, 2014.

[64] D. Eilot and R. Goldenberg, "Fully automatic model-based calcium segmentation and scoring in coronary CT angiography," *Int J Comput Assist Radiol Surg*, vol. 9, no. 4, pp. 595–608, 2014.

[65] S. Mittal, Y. Zheng, B. Georgescu, F. Vega-Higuera, S. K. Zhou, P. Meer, and D. Comaniciu, "Fast automatic detection of calcified coronary lesions in 3D cardiac CT images," in *MICCAI 2010 MLMI Workshop* (F. Wang, P. Yan, K. Suzuki, and D. Shen, eds.), vol. 6357 of *LNCS*, pp. 1–9, Springer Berlin Heidelberg, 2010.

[66] D. Dey, V. Y. Cheng, P. J. Slomka, R. Nakazato, A. Ramesh, S. Gurudevan, G. Germano, and D. S. Berman, "Automated 3-dimensional quantification of noncalcified and calcified coronary plaque from coronary CT angiography," *J Cardiovasc Comput Tomogr*, vol. 3, no. 6, pp. 372–382, 2009.

[67] S. Wesarg, M. F. Khan, and E. A. Firle, "Localizing calcifications in cardiac CT data sets using a new vessel segmentation approach," *J. Digit. Imaging*, vol. 19, no. 3, pp. 249–257, 2006.

[68] A. Schuhbaeck, Y. Otaki, S. Achenbach, C. Schneider, P. Slomka, D. S. Berman, and D. Dey, "Coronary calcium scoring from contrast coronary CT angiography using a semiautomated standardized method," *J Cardiovasc Comput Tomogr*, vol. 9, no. 5, pp. 446–453, 2015.

[69] R. A. Takx, P. A. De Jong, T. Leiner, M. Oudkerk, H. J. De Koning, C. P. Mol, M. A. Viergever, and I. Išgum, "Automated coronary artery calcification scoring in non-gated chest CT: Agreement and reliability," *PloS one*, vol. 9, no. 3, p. e91239, 2014.

[70] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *MICCAI 1998* (W. M. Wells, A. Colchester, and S. Delp, eds.), vol. 1469 of *LNCS*, pp. 130–137, Springer Berlin Heidelberg, 1998.

[71] G. Yang, P. Kitslaar, M. Frenay, A. Broersen, M. J. Boogers, J. J. Bax, J. H. Reiber, and J. Dijkstra, "Automatic centerline extraction of coronary arteries in coronary computed tomographic angiography," *Int. J. Cardiovasc. Imaging*, vol. 28, no. 4, pp. 921–933, 2012.

[72] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu, "Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features," *IEEE Trans Med Imag*, vol. 27, no. 11, pp. 1668–1681, 2008.

[73] 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 *MICCAI 2013, Part III* (K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, eds.), vol. 8151 of *LNCS*, pp. 74–81, Springer Berlin Heidelberg, 2013.

[74] S. L. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 317–328, 1997.

[75] I. Zeb, N. Abbas, K. Nasir, and M. J. Budoff, "Coronary computed tomography as a cost-effective test strategy for coronary artery disease assessment: A systematic review," *Atherosclerosis*, vol. 234, no. 2, pp. 426–435, 2014.

[76] G. Gitsioudis, W. Hosch, J. Iwan, A. Voss, E. Atsiatorme, N. P. Hofmann, S. J. Buss, S. Siebert, H. Kauczor, and E. Giannitsis, "When do we really need coronary calcium scoring prior to contrast-enhanced coronary computed tomography angiography? Analysis by age, gender and coronary risk factors," *PloS one*, vol. 9, no. 4, p. e92396, 2014.

[77] R. Tota-Maharaj, M. H. Al-Mallah, K. Nasir, W. T. Qureshi, R. Blumenthal, and M. J. Blaha, "Improving the relationship between coronary artery calcium score and coronary plaque burden: Addition of regional measures of coronary artery calcium distribution," *Atherosclerosis*, vol. 238, no. 1, pp. 126–131, 2014.

[78] E. R. Brown, R. A. Kronmal, D. A. Bluemke, A. D. Guerci, J. J. Carr, J. Goldin, and R. Detrano, "Coronary calcium coverage score: Determination, correlates, and predictive accuracy in the multi-ethnic study of atherosclerosis," *Radiology*, vol. 247, no. 3, pp. 669–675, 2008.

[79] M. G. Silverman, J. R. Harkness, R. Blankstein, M. J. Budoff, A. S. Agatston, J. J. Carr, J. A. Lima, R. S. Blumenthal, K. Nasir, and M. J. Blaha, "Baseline subclinical atherosclerosis burden and distribution are associated with frequency and mode of future coronary revascularization: Multi-ethnic study of atherosclerosis," *JACC Cardiovasc Imaging*, vol. 7, no. 5, pp. 476–486, 2014.

[80] P. H. Joshi, M. J. Blaha, R. S. Blumenthal, R. Blankstein, and K. Nasir, "What is the role of calcium scoring in the age of coronary computed tomographic angiography?," *J. Nucl. Cardiol.*, vol. 19, no. 6, pp. 1226–1235, 2012.

[81] S. Seifert, A. Barbu, S. K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu, "Hierarchical parsing and semantic navigation of full body CT data," in *Proc. SPIE Med. Imag.*, vol. 7259, p. 725902, 2009.

[82] Y. Zheng, X. Lu, B. Georgescu, A. Littmann, E. Mueller, and D. Comaniciu, "Robust object detection using marginal space learning and ranking-based multi-detector aggregation: Application to left ventricle detection in 2D MRI images," in *CVPR 2009*, pp. 1343–1350, 2009.

[83] T. Kohlberger, M. Sofka, J. Zhang, N. Birkbeck, J. Wetzl, J. Kaftan, J. Declerck, and S. K. Zhou, "Automatic multi–organ segmentation using learning-based segmentation and level set optimization," in *MICCAI 2011, Part III* (G. Fichtinger, A. Martel, and T. Peters, eds.), no. 6893 in LNCS, pp. 338–345, Springer Berlin Heidelberg, 2011.

[84] R. Cuingnet, R. Prevost, D. Lesage, L. D. Cohen, B. Mory, and R. Ardon, "Automatic detection and segmentation of kidneys in 3D CT images using random forests," in *MICCAI, Part III* (N. Ayache, H. Delingette, P. Golland, and K. Mori, eds.), vol. 7512 of *LNCS*, pp. 66–74, Springer Berlin Heidelberg, 2012.

[85] M. G. Linguraru, J. A. Pura, V. Pamulapati, and R. M. Summers, "Statistical 4D graphs for multi-organ abdominal segmentation from multiphase CT," *Med. Image Anal.*, vol. 16, no. 4, pp. 904–914, 2012.

[86] T. Okada, M. G. Linguraru, M. Hori, R. M. Summers, N. Tomiyama, and Y. Sato, "Abdominal multi-organ segmentation from CT images using conditional shape location and unsupervised intensity priors," *Med. Image Anal.*, vol. 26, no. 1, pp. 1–18, 2015.

[87] X. Zhou, S. Yoshimoto, S. Wang, H. Chen, T. Hara, R. Yokoyama, K. Masayuki, and H. Fujita, "Automated localization of solid organs in 3D CT images: A majority voting algorithm based on ensemble learning," in *MICCAI 2010 MLMI Workshop*, 2010.

[88] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med. Image Anal.*, vol. 17, no. 8, pp. 1293–1303, 2013.

[89] S. K. Zhou, "Discriminative anatomy detection: Classification vs regression," *Pattern Recognit Lett*, vol. 43, pp. 25–38, 2014.

[90] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv Neural Inf Process Syst*, pp. 1097–1105, 2012.

[91] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[92] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR 2014*, pp. 580–587, 2014.

[93] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *ICCV 2011*, pp. 1879–1886, 2011.

[94] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[95] H. S. Hecht, "Coronary artery calcium scanning: Past, present, and future," *JACC Cardiovasc Imaging*, vol. 8, no. 5, pp. 579–596, 2015.

[96] B. Messenger, D. Li, K. Nasir, J. J. Carr, R. Blankstein, and M. J. Budoff, "Coronary calcium scans and radiation exposure in the multi-ethnic study of atherosclerosis," *Int. J. Cardiovasc. Imaging*, vol. 32, no. 3, pp. 525–529, 2015.

[97] M. H. Al-Mallah, A. Aljizeeri, M. Alharthi, and A. Alsaileek, "Routine low-radiation-dose coronary computed tomography angiography," *Eur Heart J Suppl*, vol. 16, no. suppl B, pp. B12–B16, 2014.

[98] S. Voros and Z. Qian, "Agatston score tried and true: By contrast, can we quantify calcium on CTA?," *J Cardiovasc Comput Tomogr*, vol. 6, no. 1, pp. 45–47, 2012.

[99] J. M. Otton, J. T. Lønborg, D. Boshell, M. Feneley, A. Hayen, N. Sammel, K. Sesel, L. Bester, and J. McCrohon, "A method for coronary artery calcium scoring using contrast-enhanced computed tomography," *J Cardiovasc Comput Tomogr*, vol. 6, no. 1, pp. 37–44, 2012.

[100] B. Glodny, B. Helmel, T. Trieb, C. Schenk, B. Taferner, V. Unterholzner, A. Strasak, and J. Petersen, "A method for calcium quantification by means of CT coronary angiography using 64-multidetector CT: Very high correlation with Agatston and volume scores," *Eur. Radiol.*, vol. 19, no. 7, pp. 1661–1668, 2009.

[101] J. M. Wolterink, T. Leiner, B. D. De Vos, J.-L. Coatrieux, B. M. Kelm, S. Kondo, R. A. Salgado, R. Shahzad, H. Shu, M. Snoeren, *et al.*, "An evaluation of automatic coronary artery calcium scoring methods with cardiac CT using the orCaScore framework," *Med. Phys.*, vol. 43, no. 5, pp. 2361–2373, 2016.

[102] M. Teßmann, F. Vega-Higuera, B. Bischoff, J. Hausleiter, and G. Greiner, "Automatic detection and quantification of coronary calcium on 3D CT angiography data," *CSRDC*, vol. 26, no. 1, pp. 117–124, 2011.

[103] B. D. De Vos, J. M. Wolterink, P. A. De Jong, M. A. Viergever, and I. Išgum, "2D image classification for 3D anatomy localization; Employing deep convolutional neural networks," in *Proc. SPIE Med. Imag.*, vol. 9784, p. 97841Y, 2016.

[104] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *MICCAI 2013, Part II* (K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, eds.), vol. 8150 of *LNCS*, pp. 246–253, Springer Berlin Heidelberg, 2013.

[105] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *MICCAI 2014, Part I* (P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, eds.), vol. 8673 of *LNCS*, pp. 520–527, Springer International Publishing, 2014.

[106] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber, "Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation," in *Adv Neural Inf Process Syst*, pp. 2980–2988, 2015.

[107] G. L. Raff, K. M. Chinnaiyan, R. C. Cury, M. T. Garcia, H. S. Hecht, J. E. Hollander, B. O'Neil, A. J. Taylor, and U. Hoffmann, "SCCT guidelines on the use of coronary computed tomographic angiography for patients presenting with acute chest pain to the emergency department: A report of the Society of Cardiovascular Computed Tomography Guidelines Committee," *J Cardiovasc Comput Tomogr*, vol. 8, no. 4, pp. 254–271, 2014.

[108] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR 2015*, 2015.

[109] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *AISTATS 2011*, vol. 15, pp. 315–323, 2011.

[110] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Scene parsing with multiscale feature learning, purity trees, and optimal covers," in *ICML 2012*, pp. 575–582, 2012.

[111] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv preprint arXiv:1605.02688*, 2016.

[112] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS 2010*, pp. 249–256, 2010.

[113] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J Mach Learn Res*, vol. 15, no. 1, pp. 1929–1958, 2014.

[114] N. van der Bijl, R. M. Joemai, J. Geleijns, J. J. Bax, J. D. Schuijf, A. de Roos, and L. J. Kroft, "Assessment of Agatston coronary artery calcium score using contrast-enhanced CT coronary angiography," *Am. J. Roentgenol.*, vol. 195, no. 6, pp. 1299–1305, 2010.

[115] G. Funka-Lea, Y. Boykov, C. Florin, M.-P. Jolly, R. Moreau-Gobard, R. Ramaraj, and D. Rinck, "Automatic heart isolation for CT coronary visualization using graph-cuts," in *ISBI 2006*, pp. 614–617, 2006.

[116] X. Zhuang, W. Bai, J. Song, S. Zhan, X. Qian, W. Shi, Y. Lian, and D. Rueckert, "Multi-atlas whole heart segmentation of CT data using conditional entropy for atlas ranking and selection," *Med. Phys.*, vol. 42, no. 7, pp. 3822–3833, 2015.

[117] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR 2015*, pp. 3431–3440, 2015.

[118] J. Dodge, B. G. Brown, E. L. Bolson, and H. T. Dodge, "Lumen diameter of normal human coronary arteries. influence of age, sex, anatomic variation, and left ventricular hypertrophy or dilation," *Circulation*, vol. 86, no. 1, pp. 232–246, 1992.

[119] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu, "3D deep learning for efficient and robust landmark detection in volumetric data," in *MICCAI 2015, Part I* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), vol. 9349 of *LNCS*, pp. 565–572, Springer International Publishing, 2015.

[120] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans Med Imag*, vol. 35, no. 5, pp. 1160–1169, 2016.

[121] C. Hong, C. R. Becker, U. J. Schoepf, B. Ohnesorge, R. Bruening, and M. F. Reiser, "Coronary artery calcium: Absolute quantification in nonenhanced and contrast-enhanced multi–detector row CT studies," *Radiology*, vol. 223, no. 2, pp. 474–480, 2002.

[122] U. Hoffmann, U. Siebert, A. Bull-Stewart, S. Achenbach, M. Ferencik, F. Moselewski, T. J. Brady, J. M. Massaro, and C. J. O'Donnell, "Evidence for lower variability of coronary artery calcium mineral mass measurements by multi-detector computed tomography in a community-based cohort - Consequences for progression studies," *European J Radiol*, vol. 57, no. 3, pp. 396–402, 2006.

[123] J. Leipsic, S. Abbara, S. Achenbach, R. Cury, J. P. Earls, G. J. Mancini, K. Nieman, G. Pontone, and G. L. Raff, "SCCT guidelines for the interpretation and reporting of coronary CT angiography: a report of the society of cardiovascular computed tomography guidelines committee," *J Cardiovasc Comput Tomogr*, vol. 8, no. 5, pp. 342–358, 2014.

[124] D. Lesage, E. D. Angelini, I. Bloch, and G. Funka-Lea, "A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes," *Med. Image Anal.*, vol. 13, no. 6, pp. 819–845, 2009.

[125] K. Krissian, H. Bogunovic, J. Pozo, M. Villa-Uriol, and A. Frangi, "Minimally interactive knowledge-based coronary tracking in CTA using a minimal cost path," *The Insight Journal*, 2008.

[126] O. Friman, M. Hindennach, C. Kühnel, and H.-O. Peitgen, "Multiple hypothesis template tracking of small 3D vessel structures," *Med. Image Anal.*, vol. 14, no. 2, pp. 160–171, 2010.

[127] C. Zhou, H.-P. Chan, A. Chughtai, S. Patel, L. M. Hadjiiski, J. Wei, and E. A. Kazerooni, "Automated coronary artery tree extraction in coronary CT angiography using a multiscale enhancement and dynamic balloon tracking (MSCAR-DBT) method," *Comput. Med. Imaging Graph.*, vol. 36, no. 1, pp. 1–10, 2012.

[128] S. Cetin and G. Unal, "A higher-order tensor vessel tractography for segmentation of vascular structures," *IEEE Trans Med Imag*, vol. 34, no. 10, pp. 2172–2185, 2015.

[129] D. Lesage, E. D. Angelini, G. Funka-Lea, and I. Bloch, "Adaptive particle filtering for coronary artery segmentation from 3D CT angiograms," *Comput Vis Image Und*, vol. 151, pp. 29–46, 2016.

[130] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[131] M. Schneider, S. Hirsch, B. Weber, G. Székely, and B. H. Menze, "Joint 3-D vessel segmentation and centerline extraction using oblique hough forests with steerable filters," *Med. Image Anal.*, vol. 19, no. 1, pp. 220–249, 2015.

[132] A. Sironi, E. Türetken, V. Lepetit, and P. Fua, "Multiscale centerline detection," *IEEE Trans Pattern Anal Mach Intell*, vol. 38, no. 7, pp. 1327–1341, 2016.

[133] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, and G. Yang, "Hierarchical retinal blood vessel segmentation based on feature and ensemble learning," *Neurocomputing*, vol. 149, pp. 708–717, 2015.

[134] P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning for multi-task medical image segmentation in multiple modalities," in *MICCAI 2016, Part II* (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, eds.), vol. 9901 of *LNCS*, pp. 478–486, Springer International Publishing, 2016.

[135] A. Wu, Z. Xu, M. Gao, M. Buty, and D. J. Mollura, "Deep vessel tracking: A generalized probabilistic approach via deep learning," in *ISBI 2016*, pp. 1363–1367, 2016.

[136] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[137] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML 2015*, pp. 448–456, 2015.

[138] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*, 2015.

[139] X. Wang, T. Heimann, P. Lo, M. Sumkauskaite, M. Puderbach, M. de Bruijne, H. Meinzer, and I. Wegner, "Statistical tracking of tree-like tubular structures with efficient branching detection in 3D medical image data," *Phys. Med. Biol.*, vol. 57, no. 16, p. 5325, 2012.

[140] S. Dieleman, J. Schlüter, and C. R. et al., "Lasagne: First release.," Aug. 2015.

[141] B. F. Waller, C. M. Orr, J. D. Slack, C. A. Pinkerton, J. Van Tassel, and T. Peters, "Anatomy, histology, and pathology of coronary arteries: A review relevant to new interventional and imaging techniques- Part I," *Clin. Cardiol.*, vol. 15, no. 6, pp. 451–457, 1992.

[142] W. Kristanto, P. M. van Ooijen, M. C. Jansen-van der Weide, R. Vliegenthart, and M. Oudkerk, "A meta analysis and hierarchical classification of HU-based atherosclerotic plaque characterization criteria," *PloS one*, vol. 8, no. 9, p. e73460, 2013.

[143] M. Schaap, T. van Walsum, L. Neefjes, C. Metz, E. Capuano, M. de Bruijne, and W. Niessen, "Robust shape regression for supervised vessel segmentation and its application to coronary segmentation in CTA," *IEEE Trans Med Imag*, vol. 30, no. 11, pp. 1974–1986, 2011.

[144] J. M. Wolterink, T. Leiner, B. D. de Vos, R. W. van Hamersvelt, M. A. Viergever, and I. Išgum, "Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks," *Med. Image Anal.*, vol. 34, pp. 123 – 136, 2016.

[145] C. Florin, N. Paragios, and J. Williams, "Particle filters, a quasi-monte carlo solution for segmentation of coronaries," in *MICCAI 2005, Part I* (J. S. Duncan and G. Gerig, eds.), vol. 3749 of *LNCS*, pp. 246–253, Springer Berlin Heidelberg, 2005.

[146] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," *arXiv preprint arXiv:1610.10099*, 2016.

[147] A. v. d. 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," *arXiv preprint arXiv:1609.03499*, 2016.

[148] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Dilated convolutional neural networks for cardiovascular mr segmentation in congenital heart disease," in *MICCAI 2016 HVSMR Workshop* (M. A. Zuluaga, K. Bhatia, B. Kainz, M. H. Moghari, and D. F. Pace, eds.), vol. 10129 of *LNCS*, pp. 95–102, Springer International Publishing, 2017.

[149] J. Hausleiter, T. Meyer, F. Hermann, M. Hadamitzky, M. Krebs, T. C. Gerber, C. McCollough, S. Martinoff, A. Kastrati, A. Schömig, *et al.*, "Estimated radiation dose associated with cardiac CT angiography," *JAMA*, vol. 301, no. 5, pp. 500–507, 2009.

[150] J. Padgett, A. M. Biancardi, C. I. Henschke, D. Yankelevitz, and A. P. Reeves, "Local noise estimation in low-dose chest CT images," *IJCARS*, vol. 9, no. 2, pp. 221–229, 2014.

[151] L. L. Geyer, U. J. Schoepf, F. G. Meinel, J. W. Nance Jr, G. Bastarrika, J. A. Leipsic, N. S. Paul, M. Rengo, A. Laghi, and C. N. De Cecco, "State of the art: Iterative CT reconstruction techniques," *Radiology*, vol. 276, no. 2, pp. 339–357, 2015.

[152] M. J. Willemink, P. A. de Jong, T. Leiner, L. M. de Heer, R. A. Nievelstein, R. P. Budde, and A. M. Schilham, "Iterative reconstruction techniques for computed tomography part 1: Technical principles," *Eur. Radiol.*, vol. 23, no. 6, pp. 1623–1631, 2013.

[153] M. J. Willemink, T. Leiner, P. A. de Jong, L. M. de Heer, R. A. Nievelstein, A. M. Schilham, and R. P. Budde, "Iterative reconstruction techniques for computed tomography part 2: Initial results in dose reduction and image quality," *Eur. Radiol.*, vol. 23, no. 6, pp. 1632–1642, 2013.

[154] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Model Simul*, vol. 4, no. 2, pp. 490–530, 2005.

[155] Z. Li, L. Yu, J. D. Trzasko, D. S. Lake, D. J. Blezek, J. G. Fletcher, C. H. McCollough, and A. Manduca, "Adaptive nonlocal means filtering based on local noise level for CT denoising," *Med. Phys.*, vol. 41, no. 1, p. 011908, 2014. 011908.

[156] M. Green, E. M. Marom, N. Kiryati, E. Konen, and A. Mayer, "Efficient low-dose CT denoising by locally-consistent non-local means (LC-NLM)," in *MICCAI 2016, Part III* (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, eds.), vol. 9902 of *LNCS*, pp. 423–431, Springer International Publishing, 2016.

[157] A. M. Mendrik, E.-J. Vonken, A. Rutten, M. A. Viergever, and B. van Ginneken, "Noise reduction in computed tomography scans using 3-D anisotropic hybrid diffusion with continuous switch," *IEEE Trans Med Imag*, vol. 28, no. 10, pp. 1585–1594, 2009.

[158] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang, "Low-dose CT via convolutional neural network," *Biomed Opt Express*, vol. 8, no. 2, pp. 679–694, 2017.

[159] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction," *arXiv preprint arXiv:1610.09736*, 2016.

[160] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv Neural Inf Process Syst*, pp. 2672–2680, 2014.

[161] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[162] C. W. Kim and J. H. Kim, "Realistic simulation of reduced-dose CT with noise modeling and sinogram synthesis using DICOM CT images," *Med. Phys.*, vol. 41, no. 1, p. 011901, 2014.

[163] M. J. Willemink, A. M. den Harder, W. Foppen, A. M. Schilham, R. Rienks, E. M. Laufer, K. Nieman, P. A. de Jong, R. P. Budde, H. M. Nathoe, *et al.*, "Finding the optimal dose reduction and iterative reconstruction level for coronary calcium scoring," *J Cardiovasc Comput Tomogr*, vol. 10, no. 1, pp. 69–75, 2016.

[164] J. A. van Osch, M. Mouden, J. A. van Dalen, J. R. Timmer, S. Reiffers, S. Knollema, M. J. Greuter, J. P. Ottervanger, and P. L. Jager, "Influence of iterative image reconstruction on CT-based calcium score measurements," *Int. J. Cardiovasc. Imaging*, vol. 30, no. 5, pp. 961–967, 2014.

[165] C. Gebhard, M. Fiechter, T. A. Fuchs, J. R. Ghadri, B. A. Herzog, F. Kuhn, J. Stehli, E. Müller, E. Kazakauskaite, O. Gaemperli, *et al.*, "Coronary artery calcium scoring: influence of adaptive statistical iterative reconstruction using 64-MDCT," *Int. J. Cardiol.*, vol. 167, no. 6, pp. 2932–2937, 2013.

[166] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML 2013*, 2013.

[167] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *ICLR 2015 Workshops*, 2015.

[168] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," in *NIPS 2016 Workshops*, 2016.

[169] D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," *arXiv preprint arXiv:1612.05362*, 2016.

[170] N. van der Werf, M. Willemink, T. Willems, M. Greuter, and T. Leiner, "Influence of dose reduction and iterative reconstruction on CT calcium scores: A multi-manufacturer dynamic phantom study," *Int. J. Cardiovasc. Imaging*, pp. 1–16, 2017.

[171] A. M. den Harder, J. M. Wolterink, M. J. Willemink, A. M. Schilham, P. A. de Jong, R. P. Budde, H. M. Nathoe, I. Išgum, and T. Leiner, "Submillisievert coronary calcium quantification using model-based iterative reconstruction: A within-patient analysis," *Eur. J. Radiol.*, vol. 85, no. 11, pp. 2152–2159, 2016.

[172] A. Rutten, I. Išgum, and M. Prokop, "Coronary calcification: Effect of small variation of scan starting position on Agatston, volume, and mass scores," *Radiology*, vol. 246, no. 1, pp. 90–98, 2008.

[173] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.

[174] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *ICLR 2016*, 2016.

[175] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[176] Y. Han, J. Yoo, and J. C. Ye, "Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis," *arXiv preprint arXiv:1611.06391*, 2016.

[177] D. F. Pace, A. V. Dalca, T. Geva, A. J. Powell, M. H. Moghari, and P. Golland, "Interactive whole-heart segmentation in congenital heart disease," in *MICCAI 2015, Part III* (N. Navab, J. Hornegger, M. W. Wells, and F. A. Frangi, eds.), vol. 9351 of *LNCS*, pp. 80–88, Springer Heidelberg, 2015.

[178] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.

[179] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI 2015, Part III* (N. Navab, J. Hornegger, M. W. Wells, and F. A. Frangi, eds.), vol. 9351 of *LNCS*, pp. 234–241, Springer, 2015.

[180] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *ICLR 2016*, 2016.

[181] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*, 2015.

[182] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Med. Image Anal.*, vol. 31, pp. 77–87, 2016.

[183] A. de Brebisson and G. Montana, "Deep neural networks for anatomical brain segmentation," in *CVPR 2015 Workshops*, 2015.

[184] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans Pattern Anal Mach Intell*, 2016.

[185] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans Pattern Anal Mach Intell*, vol. 38, no. 1, pp. 142–158, 2016.

[186] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J Cognitive Neurosci*, vol. 19, no. 9, pp. 1498–1507, 2007.

[187] B. A. Landman, A. Ribbens, B. Lucas, C. Davatzikos, B. Avants, C. Ledig, D. Ma, D. Rueckert, D. Vandermeulen, F. Maes, *et al.*, *MICCAI 2012 workshop on multi-atlas labeling*. CreateSpace Independent Publishing Platform, 2012.

[188] B. H. van der Velden, I. Dmitriev, C. E. Loo, R. M. Pijnappel, and K. G. Gilhuijs, "Association between parenchymal enhancement of the contralateral breast in dynamic contrast-enhanced MR imaging and outcome of patients with unilateral invasive breast cancer," *Radiology*, vol. 276, no. 3, pp. 675–685, 2015.

[189] A. Gubern-Mérida, M. Kallenberg, R. Martí, and N. Karssemeijer, "Segmentation of the pectoral muscle in breast MRI using atlas-based approaches," in *MICCAI, Part II* (N. Ayache, H. Delingette, P. Golland, and K. Mori, eds.), vol. 7511 of *LNCS*, pp. 371–378, Springer, 2012.

[190] Y. Zheng, H. Tek, G. Funka-Lea, S. K. Zhou, F. Vega-Higuera, and D. Comaniciu, "Efficient detection of native and bypass coronary ostia in cardiac ct volumes: Anatomical vs. pathological structures," in *MICCAI 2011, Part III* (G. Fichtinger, A. Martel, and T. Peters, eds.), vol. 6893 of *LNCS*, pp. 403–410, Springer Berlin Heidelberg, 2011.

[191] P. Greenland, L. LaBree, S. P. Azen, T. M. Doherty, and R. C. Detrano, "Coronary artery calcium score combined with framingham score for risk prediction in asymptomatic individuals," *JAMA*, vol. 291, no. 2, pp. 210–215, 2004.

[192] B. Bischoff, C. Kantert, T. Meyer, M. Hadamitzky, S. Martinoff, A. Schömig, and J. Hausleiter, "Cardiovascular risk assessment based on the quantification of coronary calcium in contrast-enhanced coronary computed tomography angiography," *Eur Heart J Cardiovasc Imaging*, vol. 13, no. 6, pp. 468–275, 2011.

[193] H. Kirişli, M. Schaap, C. Metz, A. Dharampal, W. B. Meijboom, S. Papadopoulou, A. Dedic, K. Nieman, M. De Graaf, M. Meijs, *et al.*, "Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography," *Med. Image Anal.*, vol. 17, no. 8, pp. 859–876, 2013.

[194] J. K. Min, J. Leipsic, M. J. Pencina, D. S. Berman, B.-K. Koo, C. van Mieghem, A. Erglis, F. Y. Lin, A. M. Dunning, P. Apruzzese, *et al.*, "Diagnostic accuracy of fractional flow reserve from anatomic CT angiography," *JAMA*, vol. 308, no. 12, pp. 1237–1245, 2012.

[195] L. Pinto and A. Gupta, "Learning to push by grasping: Using multiple tasks for effective learning," *arXiv preprint arXiv:1609.09025*, 2016.

[196] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *arXiv preprint arXiv:1612.00796*, 2016.

[197] Z. Obermeyer and E. J. Emanuel, "Predicting the future - Big data, machine learning, and clinical medicine," *N. Engl. J. Med.*, vol. 375, no. 13, p. 1216, 2016.

[198] A. L. Beam and I. S. Kohane, "Translating artificial intelligence into clinical care," *JAMA*, vol. 316, no. 22, pp. 2368–2369, 2016.

[199] H. J. Kuijf, P. Moeskops, B. D. de Vos, W. H. Bouvy, J. de Bresser, G. J. Biessels, M. A. Viergever, and K. L. Vincken, "Supervised novelty detection in brain tissue classification with an application to white matter hyperintensities," in *Proc. SPIE Med. Imag.*, vol. 9784, p. 978421, 2016.

[200] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *ICML 2016*, 2016.

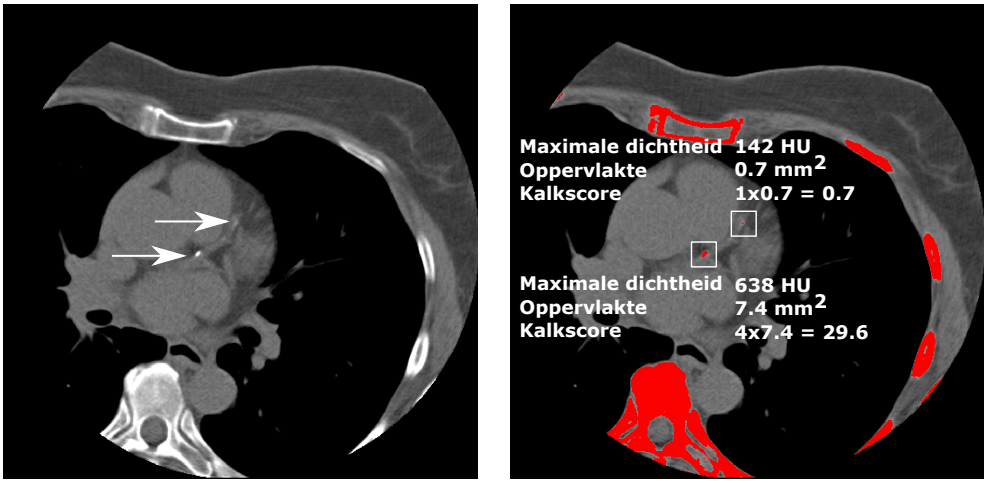
[201] M. R. Dweck, V. O. Puntmann, A. T. Vesey, Z. A. Fayad, and E. Nagel, "MR imaging of coronary arteries and plaques," *JACC Cardiovasc Imaging*, vol. 9, no. 3, pp. 306–316, 2016.

Nederlandse samenvatting

Hart- en vaatziekten

Volgens de wereldgezondheidsorganisatie zijn hart- en vaatziekten wereldwijd de voornaamste doodsoorzaak. Dit valt grotendeels te wijten aan atherosclerose, een aandoening waarbij zich in de loop der jaren plaque in de vaatwanden vormt. Atheroslerotische plaque is een ophoping van vetten, cellen en kalk. Plaque in de kransslagaders (coronairen) kan zorgen voor vernauwingen waardoor de hartspier onvoldoende zuurstofrijk bloed krijgt. Naast coronaire atherosclerose zijn er andere hart- en vaatziekten, waaronder aangeboren hartafwijkingen. Jaarlijks wordt een aanzienlijk aantal kinderen met zulke afwijkingen geboren. Als deze afwijkingen niet chirurgisch verholpen worden kan dit grote gevolgen hebben.

Bij patiënten met hart- en vaatziekten is het belangrijk een goed beeld van het hart te verkrijgen. Dat kan non-invasief door middel van computed tomography (CT) of magnetic resonance imaging (MRI). Bij CT worden met röntgenstraling projecties van het lichaam gemaakt die gecombineerd kunnen worden tot een 3D beeld. CT-beelden maken het mogelijk om vaatvernauwingen te vinden of de hoeveelheid kalk in coronaire plaques te meten. De hoeveelheid kalk in de coronairen is een sterk voorspellende maat voor hartinfarcten en overlijden. Röntgenstraling zoals gebruikt in CT is echter ioniserend, waardoor het bij overmatig gebruik schadelijke gevolgen kan hebben. Bij MRI wordt ook een 3D beeld verkregen, maar dan door gebruik te maken van de magnetische eigenschappen van waterstofatomen die overal in het lichaam aanwezig zijn. Omdat voor deze beelden geen ioniserende straling nodig is, is MRI bij uitstek geschikt voor kinderen met aangeboren hartafwijkingen.



Figuur 1: Kalkscorebepaling in hart CT zonder contrastvloeistof. De afbeelding toont het hart ter hoogte van de oorsprong van de linkerkransslagader. Twee calcificaties (≥ 130 HU) zijn aangegeven door pijlen. De maximale dichtheid in elke lesie bepaalt een wegingsfactor (1: 130-200 HU, 2: 200-300 HU, 3: 300-400 HU, 4: >400 HU). De kalkscore voor elke calcificatie wordt bepaald door deze wegingsfactor met de oppervlakte te vermenigvuldigen.

Hart CT analyse

CT-beelden van het hart kunnen met of zonder toediening van een contrastvloeistof gemaakt worden. Zo'n vloeistof maakt de bloedvaten en hartkamers beter zichtbaar in CT-beelden. Beelden zonder contrastvloeistof worden doorgaans gebruikt om de hoeveelheid coronair kalk te bepalen. Bij volledige afwezigheid van coronair kalk heeft de patiënt een laag cardiovasculair risico. Bevatten de kransslagaders wel kalk, dan is het vaak zinvol om aanvullend onderzoek te doen en eventueel preventieve maatregelen te treffen. Op basis van een hart CT-scan wordt een kalkscore bepaald die afhangt van de dichtheid en grootte van alle calcificaties (Fig. 1). Een calcificatie is een verzameling beeldpunten met een waarde boven 130 Hounsfield units (HU) in een kransslagader. De maximale dichtheid in een calcificatie bepaalt een wegingsfactor. Deze wordt met de oppervlakte van de calcificatie vermenigvuldigd om de kalkscore te berekenen. Op basis van de totale hoeveelheid kalk wordt het cardiovasculair risico van een patiënt bepaald.

Voor nauwkeurige bepaling van de kalkscore in een hart CT-scan is het belangrijk om alle calcificaties te identificeren. In de huidige klinische praktijk geeft een expert handmatig de locatie van elke calcificatie aan. Dit is tijdrovend werk dat vergemakkelijkt kan worden met behulp van automatische methoden. De methode die in Hoofdstuk 2 wordt omschreven vindt zonder tussenkomst van een expert calcificaties en bepaalt op basis van de gevonden calcificaties een kalkscore. Uit experimenten bleek dat deze methode in staat is om 93% van de patiënten in de juiste

risicogroep in te delen. Daarnaast is de methode in staat om te bepalen welke calcificaties niet met zekerheid geïdentificeerd kunnen worden. Door deze calcificaties ter controle aan een expert aan te bieden liep het percentage correct gecategoriseerde patiënten op naar 99%. Dit is vergelijkbaar met de overeenkomst tussen twee verschillende experts. Zo'n check was maar in de helft van de scans nodig en was drie keer sneller dan volledig handmatige identificatie. Het aantal fouten werd dus aanzienlijk lager, met beperkte inspanning voor de expert.

Naast de methode uit Hoofdstuk 2 zijn er in de literatuur meer technieken voorgesteld voor automatische kalkscorebepaling. Elke methode is echter ontwikkeld en geëvalueerd met andere CT-scans. Hierdoor is het moeilijk om te bepalen hoe de verschillende technieken zich tot elkaar verhouden, en wat zwakke en sterke punten in de praktijk zouden zijn. Om dit probleem op te lossen hebben we een gestandaardiseerd platform opgezet en onderzoekers uitgenodigd om hun methode toe te passen op een zorgvuldig samengestelde set scans. In dit platform worden alle resultaten op dezelfde manier geëvalueerd. In Hoofdstuk 3 wordt een evaluatie van vijf methoden binnen dit platform besproken. De resultaten laten zien dat automatische methoden goed in staat zijn om patiënten in risicogroepen in te delen. Desondanks blijft het lastig om bepaalde calcificaties goed te identificeren, bijvoorbeeld op het punt waar de kransslagader uit de aorta ontspringt.

Een standaard hart CT onderzoek bestaat uit twee scans: één zonder contrastvloeistof en één met contrastvloeistof. De eerste scan wordt gebruikt om de kalkscore te bepalen en de tweede scan wordt gebruikt om vernauwingen in de kransslagaders te identificeren. Recent onderzoek heeft aangetoond dat de kalkscore ook in scans met contrastvloeistof bepaald kan worden, waardoor de eerste scan mogelijk achterwege zou kunnen worden gelaten. Het vinden van calcificaties in scans met contrastvloeistof is echter tijdrovend. Daarom wordt in dit proefschrift een automatische methode met dit doel omschreven. Een eerste stap hierbij is het lokaliseren van het hart in de CT-scan. De methode in Hoofdstuk 4 bekijkt het beeld in drie richtingen en bepaalt voor elke richting of het hart zichtbaar is. Aan de hand hiervan wordt een 3-dimensionale doos om het hart bepaald. De methode in Hoofdstuk 5 beperkt het zoekgebied voor calcificaties vervolgens tot deze doos. De resultaten laten zien dat automatisch bepaalde kalkscores in scans met contrastvloeistof goed overeenstemmen met handmatig bepaalde kalkscores in scans zonder contrastvloeistof.

Naast de kalkscore worden hart CT-scans gebruikt om vernauwingen in de kransslagaders te identificeren. Hiervoor is het belangrijk om de loop en breedte van de kransslagaders te bepalen. Hoofdstuk 6 omschrijft een methode waarmee de loop en breedte van een volledige kransslagader op basis van één muisklik snel bepaald kan worden. De methode in Hoofdstuk 9 bepaalt voor elke beeldpunt in een hart CT-scan of het onderdeel uitmaakt van de kransslagader of niet.

Hart CT-beelden worden verkregen met röntgenstraling die bij overmatig gebruik schadelijke gevolgen zou kunnen hebben. In de afgelopen jaren is de stra-

lingsdosis waarmee CT-scans worden verkregen daarom flink naar beneden gegaan. Dat kan ervoor zorgen dat CT-scans soms veel ruis bevatten, wat diagnose moeilijker kan maken. In Hoofdstuk 7 wordt een methode omschreven die de hoeveelheid ruis in een CT-scan aanzienlijk vermindert en tegelijkertijd een realistisch ogend beeld genereert.

Hart MRI analyse

Chirurgisch ingrijpen is vaak nodig bij patiënten met aangeboren hartafwijkingen. Recent onderzoek heeft aangetoond dat een 3D model van het hart van de patiënt nuttig kan zijn bij de voorbereiding van een operatie. Zo'n model kan verkregen worden door segmentatie van een hart MRI beeld. Bij segmentatie van het hart krijgt elk punt in het beeld een label: achtergrond, hartspier of bloed. Handmatige segmentatie van een beeld is een tijdrovende taak die tot acht uur per patiënt in beslag kan nemen.

Hoofdstuk 8 omschrijft een methode waarmee een MRI beeld van het hart van patiënten met aangeboren hartafwijkingen volledig automatisch gesegmenteerd kan worden in minder dan een minuut. De resultaten tonen een sterke overeenkomst tussen handmatig en automatisch verkregen resultaten.

Machine learning

De methoden die in dit proefschrift worden omschreven, maken allemaal gebruik van machine learning. Machine learning is een aanpak voor medische beeldanalyse waarbij een computerprogramma op basis van voorbeelden leert welke beslissing er voor welk beeld genomen moet worden. Dit is anders dan regelgebaseerde systemen die aan de hand van een verzameling voorgeprogrammeerde regels handelen.

Om te leren welke beslissing er voor welke invoer moet worden genomen, heeft een machine learning methode voorbeelden nodig. Zo leert een methode voor segmentatie van hart MRI beelden aan de hand van handmatig ingetekende beelden. Omdat het veel tijd kost om beelden handmatig in te tekenen, zijn referentiesegmentaties vaak schaars. Één van de uitdagingen in machine learning onderzoek is er voor te zorgen dat er op basis van weinig data toch veel geleerd kan worden, en dat de door het systeem opgedane kennis ook nuttig is bij de analyse van nieuwe beelden.

Machine learning kan verder worden opgedeeld in conventionele machine learning en deep learning. In conventionele machine learning worden gegevens, bijvoorbeeld een beeldpunt of een beeld, omschreven met een aantal eigenschappen. Op basis van deze eigenschappen leert het algoritme onderscheid te maken tussen verschillende groepen kandidaten. In Hoofdstuk 2 worden potentiële calcificaties bijvoorbeeld omschreven aan de hand van o.a. locatie, vorm en grootte. Beslissingsbomen leren vervolgens om op basis van deze eigenschappen calcificaties te

onderscheiden. Hoewel deze beslissingsbomen leren welke combinaties van eigenschappen belangrijk zijn, blijft het definiëren van zulke eigenschappen in dit geval mensenwerk. Hierdoor wordt mogelijk niet alle informatie in het beeld benut.

Deep learning methoden gebruiken kunstmatige neurale netwerken – wiskundige modellen die gebaseerd zijn op het menselijk brein. In tegenstelling tot conventionele machine learning werken deep learning methoden niet op basis van vooraf bepaalde eigenschappen van een beeld. In plaats daarvan wordt het gehele beeld als invoer gebruikt. Een serie wiskundige operaties zet het beeld om in een uitkomst. Het netwerk leert op basis van voorbeelden hoe die operaties het beste uitgevoerd kunnen worden.

Hoofdstuk 2 laat een voorbeeld van een conventionele machine learning methode zien. Hoofdstukken 4 tot 9 laten zien hoe deep learning gebruikt kan worden voor verschillende problemen. Soms is een enkel neuraal netwerk niet in staat om een taak succesvol uit te voeren. In Hoofdstuk 5 worden daarom bijvoorbeeld twee netwerken gebruikt: één netwerk om mogelijke calcificaties te vinden en een tweede netwerk om te beslissen over de twijfelgevallen. In Hoofdstuk 7 spelen twee netwerken een spel, waarbij het eerste netwerk probeert het tweede netwerk te misleiden. Dit zorgt ervoor dat het eerste netwerk CT-beelden genereert die lastig van echte CT-beelden te onderscheiden zijn. Aan de andere kant kan een enkel neuraal netwerk juist heel veelzijdig zijn. Hoofdstuk 9 laat zien hoe één neuraal netwerk structuren kan segmenteren in hart CT, borst MRI en brein MRI.

Toekomstperspectief

Hoofdstukken 2, 3 and 5 van dit proefschrift laten zien dat de coronaire kalkscore bepaald kan worden in hart CT-scans met of zonder contrastvloeistof. Dit zou kunnen betekenen dat één scan in de klinische praktijk achterwege gelaten kan worden. Voordat het zover is, zou de methode in Hoofdstuk 5 moeten worden geëvalueerd in een groter aantal scans en gerelateerd moeten worden aan klinische uitkomstwaarden.

In dit proefschrift worden uiteenlopende methoden beschreven om op basis van machine learning technieken medische beelden te analyseren. Dit zijn vaak specifieke algoritmes, ontwikkeld voor individuele toepassingen. Deze methoden zouden echter ook gebruikt kunnen worden in andere situaties. Zo zou in de toekomst onderzocht kunnen worden of de in Hoofdstuk 6 voorgestelde methode om kransslagaders te volgen ook gebruikt kan worden om andere vaatstructuren te analyseren, bijvoorbeeld in hart MRI beelden. Daarnaast kunnen kunstmatige neurale netwerken bijzonder generiek zijn. In Hoofdstuk 9 wordt een systeem omschreven dat zowel hersenscans, hartscans als borstscans kan evalueren. In de toekomst kan dit systeem uitgebreid worden om continu nieuwe taken te leren.

Tot slot is het belangrijk hoe machine learning technieken in de klinische praktijk gebruikt gaan worden. Recent onderzoek laat zien dat de computer de mens bij

sommige klinische taken, zoals classificatie van dermatologische aandoeningen, kan verslaan. Bij andere taken heeft de computer echter nog hulp nodig van de mens en vice versa. Voor zulke toepassingen is er nog meer onderzoek nodig naar manieren om de mens en computer optimaal te laten samenwerken.