Master Thesis

TU Delft

Biomedical Imaging Group Rotterdam

Segmenting and Detecting Carotid Plaque Components in MRI

Author
Arno van Hilten
4134680

Committee
Prof. Dr. W.J. Niessen
Associate prof. M. de Bruijne
Assistant prof. J.F.P. Kooij
MSc. Z.S. Gamechi

January 18, 2018

A thesis submitted for the degree of Master of Science in Mechanical Engineering, to be defended the 29th of January 2018 at Delft University of Technology

Preface

I would like to thank all the members of the Biomedical Imaging Group Rotterdam (BIGR) for being supportive, friendly and helpful. Especially the members of the Model Based Imaging group who were always there to support me and to give me advice. Most importantly Marleen de Bruijne and Zahra Sedghi Gamechi your input was invaluable for making this thesis.

In addition I would like to thank my parents and brother for always being supportive, interested and patient.

Abstract

Segmenting and Detecting Carotid Plaque Components in MRI

Cardiovascular diseases and stroke are currently the leading causes of death worldwide. Atherosclerotic plaque is a mostly asymptotic vascular disease, but rupture of an atherosclerotic plaque in the carotid artery could lead to stroke. Automated segmentation of plaque components could help improve risk assessment by producing fast and reliable results while saving costs.

In this thesis two extensive comparisons have been made. First supervised classifiers are compared in the pixel-wise segmentation task of plaque components. In this comparison five conventional machine learning techniques and one deep learning architecture have been evaluated: linear and quadratic Bayes normal classifiers, linear logistic classifier, random forest and a U-net architecture. In the second comparison classifiers are evaluated in a detection task for their ability to learn with weakly labelled data. This is done within the multiple instance learning (MIL) framework. In addition to conventional multiple instance learning algorithms, a new MIL adaptation of the deep learning architecture, MIL U-net, is proposed and evaluated.

In the pixel-wise segmentation tasks the U-net architecture was the best overall classifier after the addition of 93 extra training patients to the original 20 training patients. A good inter-rater agreement was found for the haemorrhage class (ICC = 0.684) and the calcification class (ICC = 0.627). In the detection task the supervised methods, trained with one-sided noise, outperformed multiple instance classifiers such as MIL-Boost and the proposed MIL U-net. In this task both random forest and the linear logistic classifier obtained a fair Cohen's kappa (0.419 and 0.445 respectively) for detection of calcification per slice. The same classifiers obtained good correlation (Cohen's kappa 0.717 and 0.666 respectively) for haemorrhage detection per slice.

Contents

Chapter 1

Introduction

1.1 Atherosclerosis

Cardiovascular diseases and stroke have been the leading cause of death worldwide for the last 15 years. Ischaemic heart disease and stroke account together for 15 million deaths of the 56.4 million deaths worldwide [1]. The main cause of death, ischaemic heart disease, is a disease characterized by reduced blood supply to the heart. Ischaemic strokes (87% of the total amount of strokes) are caused by restricted blood supply to the brain. The major underlying condition for these diseases is atherosclerosis. Atherosclerotic plaque reduces blood flow in the vessels (figure 1.1) but more threatening is plaque rupture. Plaque rupture can causes blood clots which can abruptly occlude the vessel where the plaque is located or occlude smaller vessels downstream if the blood clot breaks off (embolism).

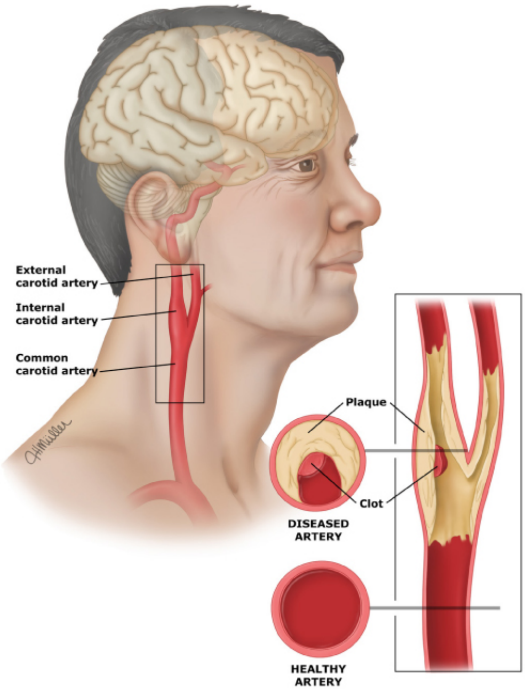


Figure 1.1: Atherosclerosis in the carotid artery. Plaques builds up inside the artery walls reducing the blood flow. If plaque ruptures a blood clot can form which could lead to stroke. Image retrieved from: https://goo.gl/a68AMn, Carotid Artery Disease & Stroke, Victorian Vascular Clinic.

1.1.1 Pathology

Atherosclerosis is an inflammatory disease in which fatty substances, cholesterol, cellular waste products, calcium and other substances build up in the inner lining of an artery, the endothelium. It is a chronic disease that remains asymptomatic for several years. Most symptoms do not occur until plaque ruptures and blockage occurs. Atherosclerosis is suspected to start when initial damaged endothelium cells triggers an inflammatory response. Macrophages enter the artery wall and start ingesting the oxidized low density lipoproteins. Eventually, after ingesting multiple oxidized low density proteins, the macrophages will die. Dead macrophages, now called foam cells, stay within the wall and release cytokines attracting more macrophages. This positive feedback system causes foam cells to accumulate, creating a possible thrombogenic lesion, often called a 'fatty streak'. In addition to these activities platelets adhere to the damaged endothelium cells producing platelet-derived growth factor which induces smooth muscle cells growth. Smooth muscle cells migrate from the tunica media and form a capsule covering the fatty streak. Smooth muscles ingest lipids and replace it by collagen, elastin and proteoglycans of the fibrous matrix [2]. These cells form a fibrous capsule around the fatty streak, the fibrous cap. If the smooth muscle cells die in this process they become foam cells like the macrophages. The fibrous cap together with the fatty streak is called an atherosclerotic plaque.

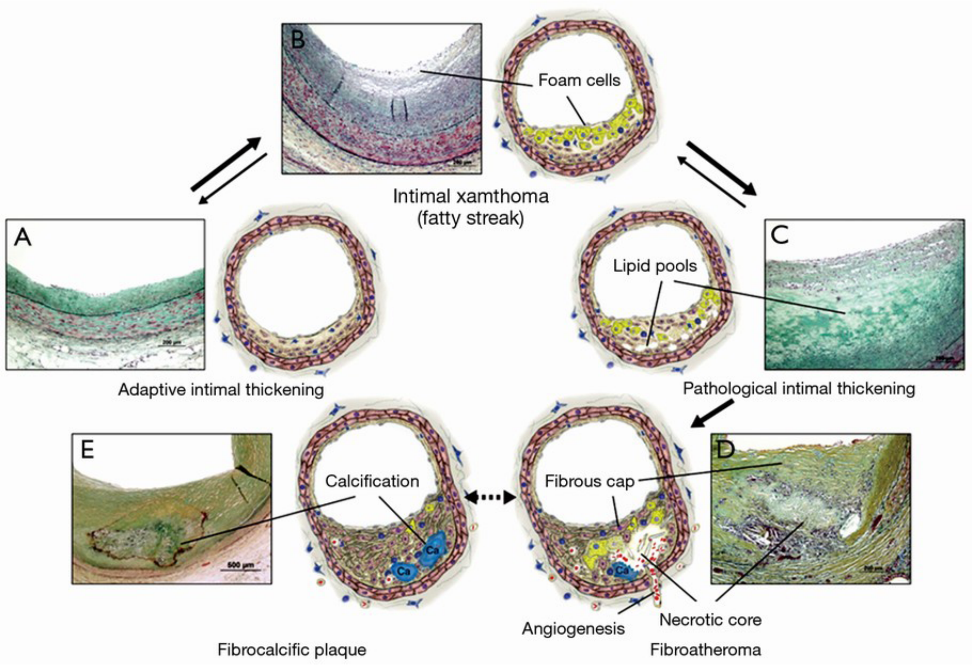


Figure 1.2: Sequence of the development of atherosclerosis as proposed by Bentzon et al. **A**: *Adaptive Intimal Thickening*, normal accumulation of smooth muscle cells in the intima. **B**: *Fatty streak*, result of accumulation of foam cells. **C**: *Pathological intimal thickening*, smaller lipid pools are seen beneath the layers of foam cells. **D**, *Fibroatheroma*, lesion with a necrotic core. **E**, *fibrocalcific plaque*, calcification of the necrotic core and surrounding tissue. The advanced lesion types **D** and **E** can evolve simiSltanously. Image retrieved from Bentzon et al. [2].

Lipid Rich Necrotic Core

Studies have shown that the necrotic core starts developing during the transition from fatty streak to fibrous plaque. The most obvious mechanism for necrotic core devolopment would be the that the necrotic core represents the debris of cells. Apoptosis and secondary necrosis of foam cells and smooth muscle cells could be the cause for necrotic core development. This could be due to hypoxia (lack of oxygen) or other cellular stresses [2], but the chemical composition of necrotic cores suggests that there may be other contributors [3]. Contributors such as the accumulation of lipids in the extracellular matrix or intraplaque haemorrhage following immature neovessel formation [4][5]. The size of the lipid rich necrotic core could indicate the risk of future surface disruptions in carotid arteries [6].

Intraplaque Haemorrhage

In an atherosclerotic plaque new microvessels may grow into the base of the plaque to supply nutients and oxygen. The new thin-walled microvessels consist of a single lining of endothelial cells on a basement membrane without the support of smooth muscle cells. In some of these vessels the endothelial lining is incomplete and as a consequence fragile and leaky [7]. The new microvessels also provide a new pathway for macrophages and other immume cells raising the possibility that it serves as a inflammatory stimulus [5]. Intra-plaque haemorrhage is common in fibrous plaques and is associated with expansion of the necrotic core and plaque instability [2][8]. Presence of intra-plaque haemorrhage is associated with approximately 5.6-fold higher risk for cerebrovascular events (such as stroke) compared to patients without intra-plaque haemorrhage [9].

Calcification

Atherosclerotic plaques frequently calcify. The calcification process can begin early during plaque development and accelerate as the plaque develops. In chronic inflammatory diseases calcification is thought to be a passive precipitation of the lesion caused by necrosis and tissue degeneration, but some evidence indicates that proteins controlling bone mineralization are also involved in the development of an atherosclerotic plaque [10][11]. Sangiorgi et al. [12] concluded that calcium is an excellent method of atherosclerotic plaque presence and that the amount of calcification correlates with the magnitude of the plaque burden in coronary arteries. Calcified plaques in the (extracranial) carotid artery are significantly less likely to be symptomatic and are more stable than plaques without calcification. Calcified plaques could thus be a marker for plaque stability [13][14][15].

Risks

Atherosclerosis is chronic disease with a long asymptomatic phase and is therfore hard to detect. Often atherosclerosis remains undetected until the first major event. Risk factores for atherosclerosis are age, gender, hypercholesterolemia, hypertension, smoking and diabetes. Local risk factors influence the the position in the vessel. Hemodynamic factors such as flow velocity and shear stress in the vessel wall are risk factors for the formation and development of plaque [16].

During progression plaque can become unstable with high chance of rupture of the fibrous cap, exposing the underlying foam cells to the blood. Platelets in the blood activate and create a blood clot blocking the vessel and preventing adequate blood supply to cells downstream which leads to the death of the cells. Blood clots may also break of from the atherosclerotic plaque site, the blood clot (embolism) is transported downstream until blocking a smaller blood vessel. Occluded coronary arteries can cause angina and myocardial infarction, occluded internal carotid arteries can cause stroke and cerebral atrophy. Not all plaques develop in unstable plaques, fibrous plaques with small or no lipid pools do have low chance of rupture and are therefore categorized as stable. Unstable or vulnerable plaques are characterized by thin fibrous caps and large lipid pools [16][17]. The composition of the plaque is considered to be an important indicator for the stability of the plaque [6][18]. Automatic segmentation of the plaque components could help large-scale population studies to investigate the relationship between plaque composition and rupture risk.

1.2 Machine Learning

What we refer to as seeing is observing patterns in the wavelength of light that falls on the retina. Brains in humans and animals have been developed incredibly well to distinguish objects, estimate distance and to recognize objects from this information. Pattern recognition is the field that studies methods to recognize patterns in images and other types of data. Computers see images as big matrices consisting of millions of pixels with different values for each intensity. An object does not always have the same intensity or shape because lightning conditions and camera angles often differ between images. It is thus hard for a computer to separate between objects and to recognize objects solely from these intensity matrices. For this reason conventional machine learning methods are not only supplied with the images but also with derived features. Features are often obtained by filtering images and by using additional information. For example filtering the images with a first order derivative could give clues where edges are located in the images. A completely different example of a useful feature could be the amount of square meters if you want to estimate the worth of a house. It is suspected that the amount of square meters does have a positive correlation with the house worth. From these examples it is immediately clear that features are often domain dependent, require some domain knowledge and that more than one feature is necessary to learn a good model. After all a large house in the countryside is worth less than a house with the same amount of square meters in the middle of a city. If an expert estimates the worth as well, this information could be used as feedback during training or to evaluate the performance of the algorithm. This experts opinion is assumed to be true and called the ground truth. Algorithm that use this truth to learn, try to mimic the experts behaviour.

Machine learning algorithms are often used to approach computer vision and pattern recognition problems. In the field of machine learning computers can be taught to 'see' by showing examples (often with features) during the training phase. If each example has a ground truth during the training phase the type of learning is called supervised learning. During training a machine learning algorithm tries to use the examples together with the ground truth labels to constructs a mapping for unlabelled instances to discrete classes in a classification problem, or a mapping to a continuous domain for regression. Estimating the house worth is thus a regression problems while categorizing medical images as healthy or unhealthy is an example of a classification problem. It is important that the mapping is able to generalize, classifying previously unseen examples correctly. To go back to the house example: the algorithm should be able to give a good estimation of the worth of different types of houses, thus also for houses that were not in the training set. To measure how well a algorithm generalizes the data is often divided in three sets: a training set to train the algorithm, a validation set to tune the parameters and a test set to get an unbiased estimate of the algorithms performance.

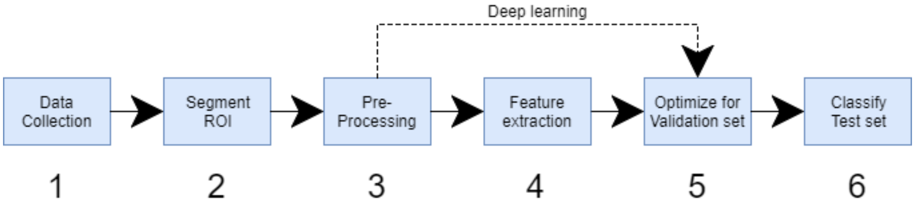
1.2.1 Types of Machine learning

Most of what humans and animals learn is unsupervised. By observing the world and discovering structure in it humans and animals learn. In general humans do not need to learn from thousends of examples as a cumputer does in a supervised setting. After a few examples humans are able to differentiate and able to generalize the learned concept. LeCun, Bengio & Hinton (2015) [19] expect that for this reason unsupervised learning will become far more important in the long term.

In the supervised framework every example in the training set has a label given by an expert (the ground truth). In the unsupervised learning framework the training set is unlabelled. Semi-supervised learning combines both: the training set contains labelled and unlabelled data. Unsupervised and semi-supervised learning methods are still able to learn from unlabelled data by inferring patterns and anomalies in the data. In this thesis supervised learning and another type of machine learning will be explored, Multiple Instance Learning (MIL). Multiple instance learning (MIL) can reduce the need for costly annotations in segmentation tasks by weakening the required degree of supervision. In the multiple instance framework a label per vessel, image or per patient can be used to train classifiers where in the supervised framework a label per pixel would have been required.

1.2.2 Procedure

The computer vision problem in this thesis can be divided in the following steps:



1. Data collection
2. Segmentation of region of interest
3. Pre-processing
4. Feature extraction
5. Optimizing for validation set
6. Classification of the test set

The proces starts with acquisition of the data, from this data the region of interest is segmented and the data is pre-processed. Pre-processing often includes normalization and outlier removal. After pre-processing the data is divided in three sets: a training set, a validation set and a test set. Features are extracted and selected to describe the differences in classes. Feature selection reduces the dimensionality by removing redundant or irrelevant features. During the classification step an appropriate classifier (machine learning algorithm) is chosen and trained with the training set. There is no single algorithm that is better than all the others on all the problems as stated by the 'No free lunch' theorem by Wolpert [20]. Following this theorem multiple algorithms and parameter settings are tested. The classifiers and features are optimized for the validation set. The results on the validation do not reflect how well the algorithm performs on previous unseen data. After all the validation set is used to tune the parameters of the learning algorithm. Although the validation set was not included in the training set the results may have been influenced by choosing the parameters based on the results of the validation set. Therefore the test set is only used once, to obtain a final unbiased estimate of the performance of the tuned algorithm.

1.3 Objectives

The algorithms in this thesis are trained to make a pixel-wise segmentation of the different plaque components (fibrous tissue, lipid rich necrotic core and haemorrhage) in the carotid artery. If succesful, the trained algorithm could be used in research to gain a better understanding of the influence of plaque components on the stability of a plaque. One of such studies within the Rotterdam Study is the study of van Bouwhuijsen et al. [21]. Their definitions for the assessment of the different plaque components will be used in this thesis. The criteria for locating the different plaque components in MRI data are:

"*Haemorrhage is defined as a hyper-intense (brighter) region on 3D-FS-Coronal. Calcification was defined as the presence of a (darker) hypo-intense region in all of the modalities. Lipid Rich Necrotic core was defined as a hypo-intense region, not classified as intra-plaque haemorrhage or calcification, in the plaque on PDw-FSE or PDw-EPI and T2w-EPI images, or a region of relative signal intensity drop in the T2w-EPI images compared with the PDw-EPI images.*"

van Bouwhuijsen et al.[21]

The remaining pixels between the vessel wall and lumen not classified as one of the other components (haemorrhage, calcification or lipid rich necrotic core) are defined as fibrous tissue.

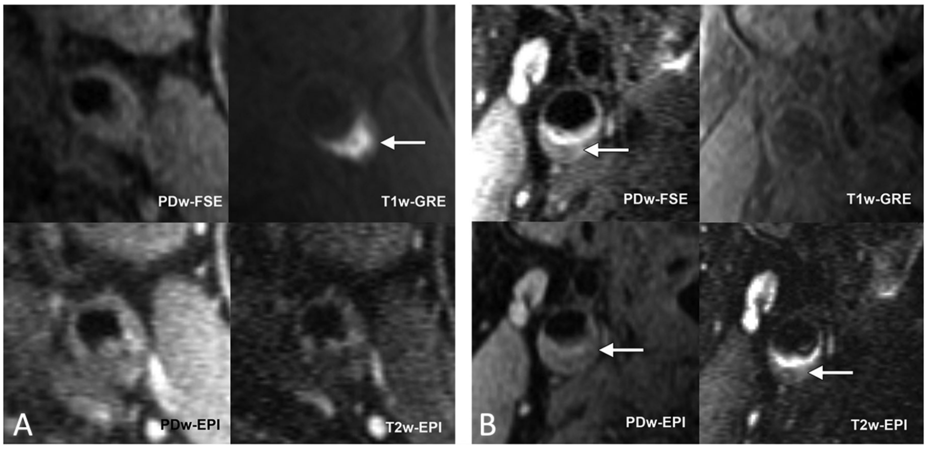


Figure 1.3: Samples of the ErgoCar study. Figure **A** shows an example of a vulnerable plaque with haemorrhage characterized by the high signal intensity (bright) in the T1-weighted gradient-echo image (T1w-GRE) annotated with the white arrow. In figure **B** white arrows denote the presence of lipid core in the relevant sequences. There is not a high signal intensity in the carotid plaque on the T1-weighted gradient-echo image. In combination with a region of low signal intensity in the plaque burden on both PDw-EPI and T2w-EPI images a lipid core can be identified. Images retrieved from Van den Bouwhuijsen et al. [22]

The main objective is to train a classifier in a way that it learns to find and segment the different plaque components as defined by van Bouwhuijsen et al. [21]. To train the classifier different types of annotations can be used. In this thesis manually created pixel-wise annotations will be used for supervised learning. Gaining pixel-wise annotations for supervised learning is labour intensive and expensive. Therefore other possibilities will be explored to reduce the labelling effort. One of the explored possibilities is to only gather manual annotations for the plaque components and combine these with automatic extracted vessel segmentations. Reducing the labelling effort to only the plaque components. The final explored possibility is to train the algorithms with 'weak' labels that only contain information whether a component is present or not. This type of information is fast to obtain and often already present (e.g. patient records, diagnosis). Within the fully supervised learning framework it is not possible to train a classifier on this type of annotations. Therefore multiple instance learning algorithms will be used to train with these type of labels.

Lately deep learning architectures have obtained state-of-the-art results in several machine learning competitions [23][24][25] and have gained more popularity in medical imaging. The increased popularity of deep learning methods can be attributed to advances in hardware making it possible to train deep neural networks in a reasonable amount of time. The methods are also quite flexible and can learn with different types of annotations with only minor adaptations, opening the possibility to train with a mixture of annotation types. This property of deep learning networks will be explored in the final experiments training with a combination of weak and strong labels in order to find alternative ways to reduce labelling effort.

The main objectives are:

1. Train a supervised algorithm to produce state of the art classification for the four different plaque components (i.e. fibrous tissue, lipid rich necrotic core, calcification and haemorrhage).
2. Investigate whether vessel-wise annotated data is a feasible annotation type for training an algorithm to detect plaque components.
3. Investigate whether a combination of vessel-wise annotated data and pixel-wise annotated data could lead to better performance in the detection task of plaque components.

Chapter 2

Materials & Pre-Processing

In this chapter all the steps followed for creating the data set will be discussed. The data used is collected within the Rotterdam Study. This study is discussed followed by data acquisition, pre-processing, elaboration on the different training sets and finally inter-observer variability .

2.1 Rotterdam Study

The Rotterdam Elderly Study is a prospective cohort study in the Ommoord district in the city of Rotterdam, the Netherlands. The main objectives of this study is to invatigate the risk factors of cardiovascular, neurological, ophthalmological and endocrine diseases in the elderly. In the year 1989 persons living in the district Ommoord from 55 years of age or over were invited to participate in the study. In 2006 this range was extended to include patients from age 45 or over. This resulted 14,926 participants at the end of 2008. All the participants are repeatedly examined every 3-4 years [26].

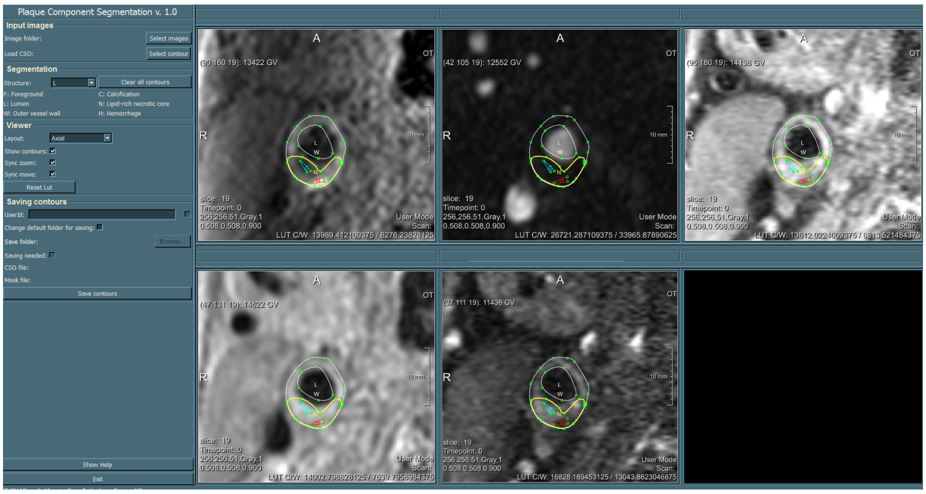
From October 2007 Magnetic Resonance Images (MRI) of the carotid artery were included in the Rotterdam Study. Subjects with carotid wall thickening bigger or equal than 2.5 mm in ultrasound are selected for MRI scans. In this thesis a subset from one scan round with a total of 685 subjects was used.

2.1.1 Carotid MRI in the Rotterdam Study

The Magnetic Resonance Images (MRI) of the carotid artery were obtained with a 1.5 Tesla scanner (GE Healthcare, Milwaukee, WI, USA) with a bilateral phased-array surface coil (Machnet, Eelde, The Netherlands). The bifurcation points of the common carotid arteries were localized and several scans were made on the same location around the bifurcation points. By using different parameters for each scan different sequences are obtained. Depending on the properties of tissue contrast with surrounding tissue can differ from one sequence to another. For example the proton density weighted fast spin echo black blood sequence (PDw-FSE-BB) shows high contrast between the blood in the lumen and the surrounding tissue. Aside from the PDw-FSE-BB four other sequences are made: a proton density weighted echo planar imaging sequence (PDw-EPI), a T2w-EPI sequence and a 3D-T1w-gradient echo (GRE) sequence. The 3D-T1w-gradient is also referred to as the 3D-FS-Coronal. The final sequence is the 3D phase-contrast MR angiography, a sequence mainly used to locate the lumen. An example of each sequence can be seen in figure 2.1.

2.2 Data Pre-Processing

The images are pre-processed with the same methods as in Arias et al.[27]. N4 Bias field correction was applied to all the sequences with default parameters. Initially the phase contrast sequence is registered to the black blood sequence in a 3D rigid registration with mutual information as similarity metric. Subsequently the sequences were registered non-rigidly to the blackblood sequence using the same parameters as in van 't Klooster et al. [28]. In this step a circulair registration mask covering the whole vessel was used . Within the same mask the intensity values between the 5th and 95th percentile are normalized by linear intensity normalization. In the final pre-processing step the images are cropped 22.5 mm below and 6.3 mm above the bifurcation point of the carotid artery[27].



Plaque Component Segmentation v. 1.0

Input images

Image folder:

Load CSO:

Segmentation

Structure: L

F: Foreground

L: Lumen

W: Outer vessel wall

C: Calcification

N: Lipid-rich necrotic core

H: Hemorrhage

Viewer

Layout: Axial

Show contours: ☑

Sync zoom: ☑

Sync move: ☑

Saving contours

UserId: ☑

Change default folder for saving: □

Save folder:

Saving needed: ☑

CSO file:

Mask file:

Figure 2.1: *Plaque Component Segmentation v. 1.0*: the annotation tool used for annotating the 90 MR images. The five different modalities are seen in this image. Top row left to right: 3D-T1w-GRE sequence, the phase contrast sequence and the PDW-EPI sequence. Bottom row: PDW-FSE-BB (black blood) sequence, and the T2W-EPI sequence. The annotations made in this tool are overlaid: the vessel outline in white, yellow for lipid rich necrotic core, blue for calcification and red for haemorrhage

2.2.1 Manual Ground Truth

An observer with several years of experience with carotid MRI data manually annotated the pre-processed MR images according to earlier discussed criteria from van Bouwhuijsen et al. [21]. Figure 2.1 shows the annotation tool with the pre-processed data and the annotations made with the tool. 90 patients were manually annotated for the components and the complete vessel. In the figure it can be seen that all the MRI sequence and annotations do align and share the same coordinate system. The annotations tell for each pixel-coordinate to which class the pixel on this coordinate belongs. It is a pixel-wise annotation. Since features are extracted per sequence it is of great importance that the sequences are correctly aligned during pre-processing in the registration step. During the project problems with the alignment of the sequences were noticed and the data was manually reviewed for errors. 39 Patients were excluded for not having any component present or due to registration errors. The remaining 51 patients were divided over a training, validation and test set. 20 patients were assigned to the training set, 10 to the validation and 21 to the test-set. This decision was based on learning curves in earlier projects and because the training set will be extended (next section). The patients were assigned to the sets based on the distribution of the plaque components. Fibrous tissue is not taken in account for the division of the patients over the sets. In this highly imbalanced data set, fibrous tissue is healthy tissue and the background class. Per component the mean amount of pixels per patient belonging to that component was calculated and this was multiplied with the set-size to estimate an optimal distribution per set. The patients were randomly assigned to a set and the error margin with respect to the optimal distribution was measured. The allowed error margin with the optimal distribution was slowly increased until a distribution of the patients was found within the allowed error margin. The final distribution can be found in figure 2.2.

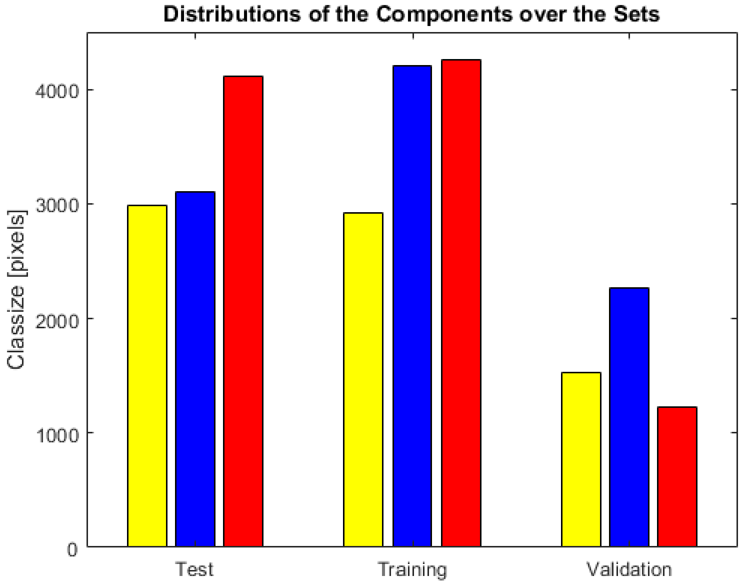


Figure 2.2: The distribution of the components over the sets. Yellow is lipid rich necrotic core, blue is calcification and red haemorrhage. The validation set is roughly half the size of the training and test set. The background class (fibrous tissue) is not included and has a magnitude of roughly 400 000 pixels for the training and test set.

2.2.2 Automatic Vessel Ground Truth

Additionally 586 patients were annotated by the same observer for the three plaque components. To include these patients in the data set it is required that the annotations contain the vessel. It is thus necessary to obtain vessel segmentations to complete these annotations. An U-net architecture was trained on the manual annotations to obtain these automatic vessel segmentations. There was no time for thorough optimization and the first results compared infavorable when compared with the segmentations obtained by Arias et al. [27]. More information on the U-net for vessel segmentation can be found in Appendix A).

The segmentation made by the algorithm of Arias et al. [27] were chosen to complete the annotations. To make sure most of the manual annotations of the plaque components were preserved and because the algorithm is suspected to under-segment the vessels in cases with large plaques, the resulting vessel-segmentations were dilated as can be seen in figure 2.3.

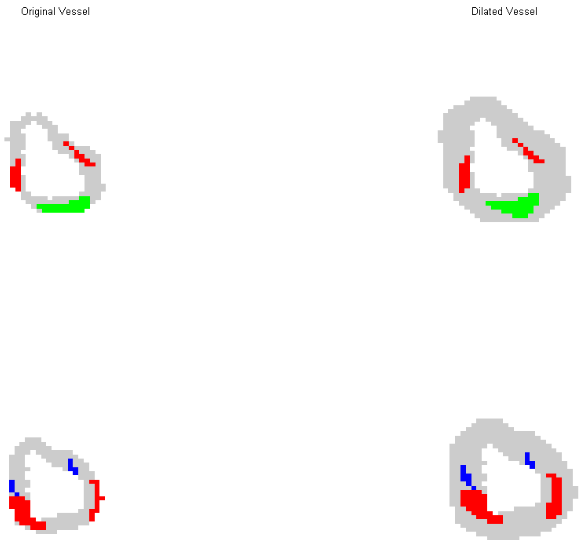


Figure 2.3: *Left* the original vessel segmentation by Arias et al. [27]. *Right* the dilated segmentation. With in red the annotations for haemorrhage, in blue the annotations for calcification and in green the annotations for lipid rich necrotic core.

Not for every patient vessel segmentations and other required information was available. A total of 109 patients had missing information such as missing sequences, annotations, vessel-segmentations or necessary meta-data. The remaining 477 patients still had errors due to registration problems. In attempt to reduce these problems, the patients were sorted by the largest translation during the rigid registration of the T1w-GRE (3D-FS-Coronal) sequence to the PDw-FSE-BB sequence (a pre-processing step). During rigid registrations small translations are expected and large translations could indicate registration problems. While inspecting the patients with largest transformations, registration errors were found as can be seen in figure 2.4. A cut-off threshold of 2 mm was arbitrarily chosen leaving 231 eligible patients. From this set 93 patients were randomly selected and added to the training set. These 93 patients are only added to the training set and not to the validation or test set because these patient annotations with automatic vessel segmentations are less reliable and contain more noise than fully manual annotated images. For reliable evaluation only manual annotations are used in the validation and test set. These 93 patients could still help improve training: for all algorithms it is tested if the addition of these 93 patients with the automatic vessel ground truth is beneficial.

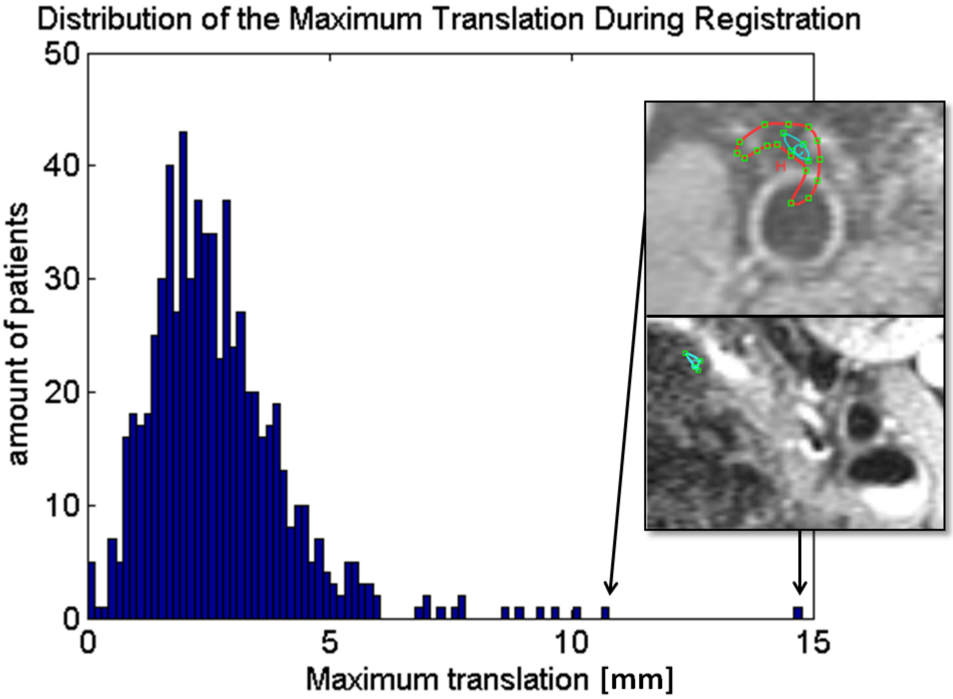


Figure 2.4: Distribution of the maximum translation during rigid registration. Examples of errors in alignment between the sequence and the annotations are shown. The overlaid annotations are not within the vessel in the shown sequences. Large translations during rigid registration seems a good indicator to find these registration errors. Patients with the largest maximum translations all had registration problems.

2.2.3 Observer Variability

By evaluating different algorithms on the same data set the performance of the classifier can be measured and compared to other algorithms. By evaluating the algorithm's performance on the ground truth the algorithm is evaluated on its ability to match the performance of a single observer. It is important to keep in mind that the annotations of a single observer is not the absolute truth. Human performance is not constant over time and two experts can have a slightly different interpretation. Inter-observer variability is the amount of variation between the results obtained by multiple observers on the same material. The agreement between observers can be given in percentages but this would not consider the possibility of agreement occurring by chance. Therefore the agreement is measured with Cohen's kappa coefficient (see section 3.4) which takes in account the possibility of the agreement occurring by chance.

The patients with the pixel-wise annotations of the first observer were a subset of a set used in a second study by Selwanesse et al. [29]. In this study another observer identified for 1663 patients solely the presence of the different plaque components per vessel. To get the agreement between the observers the annotations are converted to labels for the presence of plaque components per patient. The intersection of the sets consist of 352 patients. For these 352 patients the Cohen's kappa and observed agreement (absolute agreement in % between observers) is calculated for each component.

	Cohen's kappa (n=352) oberver 1 vs observer 2			Cohen's kappa (n=60) observer 2 vs observer 3		
	Observed agreement	Cohen's kappa	Kappa error	Observed agreement	Cohen's kappa	Kappa error
Lipid	0.5767	0.1356	0.0538	0.7167	0.4420	0.1146
Calcification	0.9744	0.2966	0.2314	0.9333	0.8537	0.0707
Haemorrhage	0.8722	0.7430	0.0358	0.9167	0.7727	0.0973

Table 2.1: Cohen's kappa for the presences of plaque components in patients. The comparison between observer 1 and observer 2 consists of 352 patients. The comparison between observer 2 and observer 3 compared 60 patients. The patient-wise observations from observer 1 are extracted from pixel-wise annotations, wile the observations of observer 2 and 3 are extracted from vessel-wise annotations.

In the first comparison between observers 352 patients were annotated by both observer 1 and 2. The unweighed Cohen's kappa for haemorrhage is substantial with a high observed agreement. The Cohen's kappa for the other components are worse. The Cohen's kappa for calcification has a value of roughly 0.29 which would be a fair agreement. The observed agreement for calcification was very high. The low kappa score is thus a result of a high random agreement of 96.36 %. Nearly all patients included in the comparison had a calcified plaque. The lowest Cohen's kappa was obtained for lipid rich necrotic core. The observed agreement for this class is fairly low with only 57.67 % agreement. It must be noted that while the annotations were acquired on the same data, the observers had different procedures. The first observer made pixel-wise annotations and has probably seen the slices for a longer period of time. These pixel-wise annotations were then transformed to patient level annotations. The observations of observer 2 are extracted from vessel-wise annotations and observer 2 has probably inspected the individual slices less extensively.

In the second comparison the observers worked under more even conditions. An observer in training identified plaque components with the same procedure as observer 2. Observer 3 identified the presence of plaque components in 60 patients from the set of the second observer. In this smaller comparison Cohen's kappa for haemorrhage was found to have a similar value as the kappa in the first comparison. The kappa values found for lipid rich necrotic core and calcification are substantially higher than in the first comparison. The difference could be explained by the high random agreement in the first comparison and the more even procedure.

Although the values found for Cohen's kappa vary between the comparisons a trend can still be seen. Haemorrhage was easier to identify for the observers followed by calcification. Lipid rich necrotic core is the most difficult component to identify consistently.

Chapter 3

Methods

3.1 Overview

The thesis is two part. The first part consists of a comparison between several supervised learning algorithms, this includes conventional supervised learning algorithms such as the linear and quadratic Bayes, the logistic classifier and random forest but also a deep learning architecture: U-net. In the second part of this thesis multiple instance learning will be explored. Conventional multiple instance learning methods will be compared to one deep multiple instance learning method proposed in this thesis: MIL U-net. Multiple instance algorithms are often based on supervised algorithms. This allows us to evaluate and compare the behaviour between the two frameworks. In both frameworks the potential benefit of deep learning compared to conventional machine learning algorithms will be evaluated.

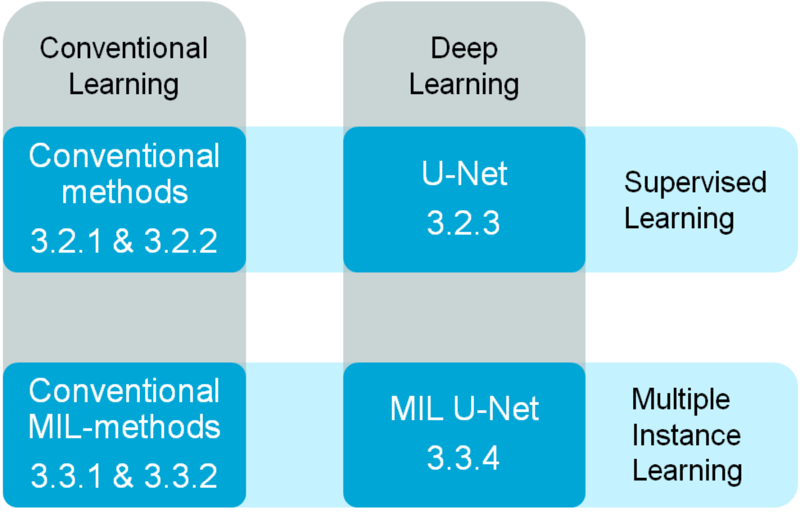


Figure 3.1: Graphical overview of algorithms in this thesis and in which section they are introduced.

3.2 Supervised learning

The majority of practical machine learning applications are made by using the supervised learning framework. In a supervised setting an algorithm is trained with a training set in which every example has a ground truth. Since the goal is to perform pixel-wise segmentation, every pixel is an example that needs to be evaluated by the algorithm. It is thus also necessary to have a ground truth for each pixel (pixel-wise annotations). There are many supervised learning methods such as: Linear Discriminant Analysis [30], Support Vector Machines [31], Random Forest [32], Neural Networks [33]. In recent years a new class of algorithms became increasingly popular: Convolutional Neural Networks [34]. Past years deep convolutional networks have achieved state of the art performances in various computer vision task such as classification, detection and segmentation tasks [23][24][25].

In this thesis conventional methods, using handcrafted features, are compared to a deep learning architecture. The features used by the conventional methods are carefully handcrafted features used in earlier work for plaque component segmentation by van Engelen et al. [35] with the addition of some texture based features. Hand-crafted features allow to insert domain-expertise, features can be based on manual procedures for segmenting or diagnosing patients. At the other hand domain expertise is often expensive and manually crafting features costs time. Deep learning architectures discover their own features by generating feature maps by using convolution operations.

For the conventionial methords only the features from the region between the lumen and vessel wall are evaluated. In total 124 features were extracted by taking the features from van Engelen et al. [35] on multiple scales and by extending it with texture based features. Per sequence the following features were obtained: the normalized MRI intensity values, scale space features such as gradient magnitude, Laplacian on three different scales, texture features such as: local range, local entropy, local standard deviation and local binary patterns. From the vessel segmentations the distance to the vessel wall and lumen were calculated as well as the multiplication between these two. Deep learning methods do have their own feature generating mechanism. For the deep learning methods the centrelines extracted in the method described in Arias et al. [27] were used to crop the images to size [96 x 96 pixels] around the centreline. The cropped images together with the vessel segmentation are used as an input for the deep learning architecture. The deep learning architecture will use these input images to create its own feature maps.

3.2.1 Random Forest

Random forest will be used as proposed by Leo Breiman [32]. The implementation [36] used is a Matlab port of the random forest from the R-source by Andy Liaw et al. [37]. This code was used in combination with the PRtools toolbox [38] in Matlab.

A random forest consists of multiple decision trees. A random forest is thus an ensemble of decision tree classifiers where each tree casts a unit vote for the most popular class for its input. Decision trees tend to classify with a low bias but do have a very high variance, they tend to easily over-fit on the training set. In a random forest the variance is reduced by growing each decision trees on a different bootstrap sample of the data set. This reduces the correlation between the trees and reduces the generalization error.

After selecting a bootstrap sample from the training data a decision tree is grown. To decrease the correlation between trees further not all features are evaluated. At each node random features are drawn and the best possible split with these features is chosen. In each bootstrap roughly two-third of the training set is used. The remaining samples not in the current bootstrap sample are the out-of-bag samples which are used to give an ongoing estimation of the generalization error. The out-of-bag error is acquired by aggregating all the votes grown on out-of-bag data. In addition the random forest gives one other unique estimate: feature importance. After constructing each tree each feature is once permuted while all other features remain unchanged. The change in the out-of-bag error gives compared to the not permuted out-of-bag error averaged over all trees gives an estimate the importance of this feature.

Random Forest has the advantages of being inherently multi-class and is relatively robust to outliers and noise [32]. Another advantage is the aforementioned internal estimate for the error (out-of-bag error) and feature importance which can be used for feature selection. However random forest do need a great amount of data and training can take considerable amount of time compared to other conventional machine learning algorithms such as the linear and quadratic Bayes.

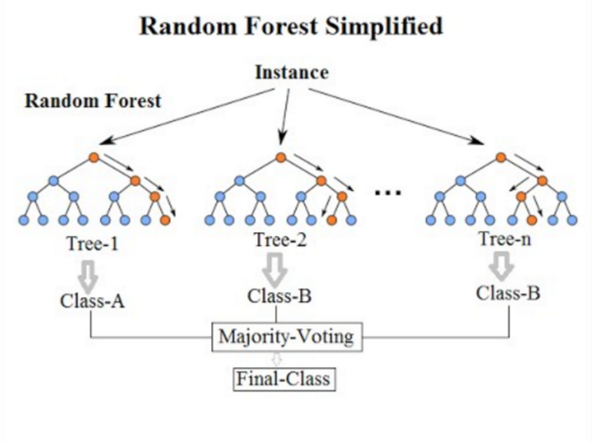


Figure 3.2: Illustration of the working principle of a random forest. Each decision tree in the ensemble casts a vote. The class with the majority votes is the prediction. Image retrieved from TIBCO Community, https://goo.gl/Y7vsKH

3.2.2 Conventional Baseline Classifiers

Other conventional machine learning algorithms will be included as additional baselines. Similar to [39] the LDC, QDC classifiers from the PRtools toolbox [38] will be used. By assuming normal densities and a common covariance matrix a linear model is created. Using a decision rule that picks the maximum posterior probability the Linear Bayes Normal Classifier (LDC) is obtained. If the assumption of the equal covariance matrix is not made the quadratic terms do not cancel and the a Quadratic Bayes Normal Classifier (QDC) is obtained. The QDC has quadratic decision boundaries and it needs to estimate the covariance matrix since this term does not cancel out anymore. The logistic linear classifier is the final baseline added from the PRtools toolbox [38] it works by maximizing the likelihood criterion using the logistic (sigmoid) function.

3.2.3 U-Net

Convolutional Neural Networks (CNN) became more viable due to the increase in computational power of GPUs in recent years. Deep learning methods such as convolution neural networks can process data in its raw form without feature extraction and thus without domain expertise. The features are replaced by generated feature maps in the convolution layers of a CNN. In most applications a window of inputs from the previous layer is chosen, called the local receptive field. This window is moved over the whole input image, convolving the inputs from the sliding window with a filter. This filter acts as a feature detector, the same weight and bias is used for all the inputs to extract a feature. Often multiple filters are learned, each extracting different features. The output matrix is thus formed by the convolving the sliding window and the filters. These matrices are called feature maps and do replace the hand-crafted features in conventional machine learning. Pooling is done to reduce the dimension of the feature maps. A pooling layer divides the input layer in regions from which only the maximum or average is passed to the next layer. Fully connected layers, every neuron connected with the previous layer, are used to get predictions. The multilayer architecture can be trained by simple stochastic gradient descent by computing gradients using backpropogation [19][40].

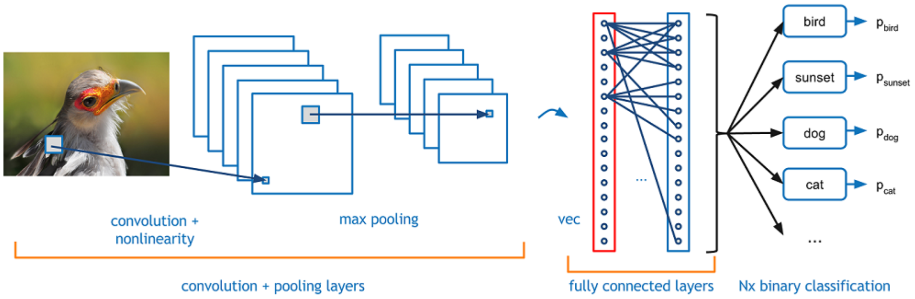


Figure 3.3: Overview of a neural network with convolution, pooling and fully connected layers. Image retrieved from: https://goo.gl/9iPjDT. Deshpande A,. (2016). Cover example of CNN.

The U-net architecture [41] is a fully convolutional neural network architecture designed to produce pixel-wise output predictions. Inspired by Ciresan [42] and Long et al. [43] the authors build a more elegant architecture and named it the U-net architecture. The name is derived form its characteristic U-shape which can be seen in figure 3.4. The down-sampling path (the first half of the U-shape) consists of a usual contracting network where convolutions are alternated with max pooling layers to down-sample the image. In order to gain high resolution output the up sampling path is added (the second part of the U-shape) which contains up-sampling followed by convolution layers. To enhance the output predictions, earlier high resolution feature maps are copied and concatenated with the up-sampled (coarser) feature maps. This allows the successive convolution layers to construct more precise output based on the coarse and fine feature maps. In contrary to the FCN-8 architecture [43] the U-net architecture has also a large number of feature maps in the up-sampling path creating an up-sampling path symmetric to the down-sampling path.

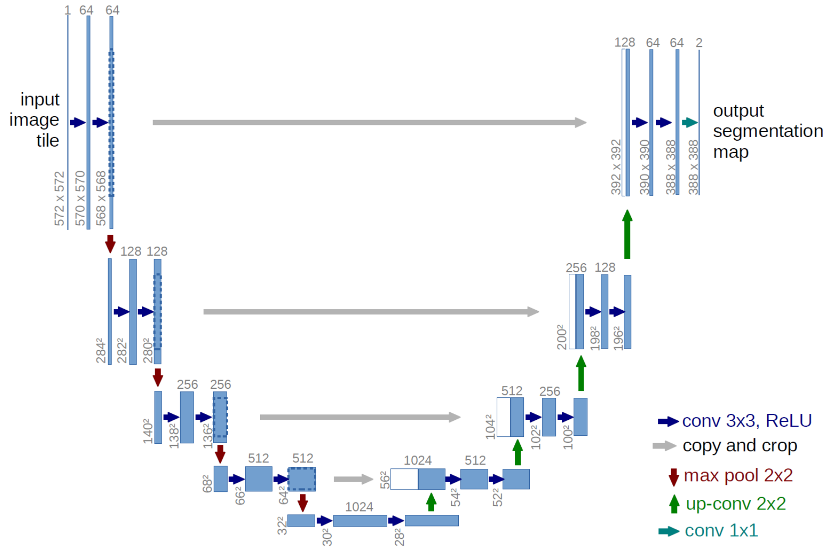


Figure 3.4: U-net architecture as used in the original paper Ronneberger et al. [41]. The eventual structure of the U-net can be seen as a hyper-parameter. The depth of the network, batch size and number of features maps and regularization parameters are optimized on the validation set (Section 4.2.2).

Adaptations to U-net for Carotid Plaque Component Segmentation

Several adjustments have been made to the standard U-net [41] to fit the segmentation task in this thesis better. A specific loss function has been chosen to fit the problem. The weights in the CNN are updated using the loss function. For multi-class classification problems categorical cross-entropy is a common choice. Since there is a class imbalance, a weight factor has been added to give the minority classes more influence. This function is normalized with respect to each slice, resulting in the following loss function:

$$\mathcal{L} = -\frac{\sum_j t_{ij} w_c \log(p_{ij})}{\sum_j t_{ij} w_c} \tag{3.1}$$

Where w_c is the weight matrix: a diagonal matrix with the weights for each class on the diagonal. t_{ij} and p_{ij} are tensors containing the target (ground truth) labels and the predicted probabilities per pixel respectively. The U-net is implemented using the python framework Keras in combination with a Theano backend. To speed up training a GPU (Nvidia geforce GTX 1070) is used in combination with CUDA. The labels are converted to the categorical format such that only the pixels within the region between the vessel wall and lumen contributed to the loss (see table 3.1). In the second channel a binary mask of the region between the vessel wall and lumen is supplied with the intend to let the network learn the region of interest and to speed up training. The loss is optimized by using the adaptive learning rate method: ADADELTA [44].

	F	**L**	**C**	**H**
Pixels in the Image	1	0	0	0
	1	0	0	0
	..	..	..	..
	0	0	1	0
	0	0	0	1
	..	..	..	..
	0	0	0	0
	0	0	0	0
	..	..	..	..
	1	0	0	0
	1	0	0	0

Table 3.1: Example of the categorical labels as used in the loss function. Each pixel can only belong to one class, in the column of that class the value is one. In the predictions this is enforced by the last layer in the network: a softmax activation layer over the rows. The row denoting calcification has a blue font and haemorrhage red. Pixels not in the region between the vessel wall and lumen belong to the background and are thus zero in each class. These pixels do not contribute to the loss.

3.2.4 Related Work in Plaque Component Segmentation

MRI has shown great potential for non-invasive, in vivo, characterization of plaque components [45][46][47] and has been used in related population studies by van Bouwhuijen [21] and Selwanesse et al. [48]. Automatic segmentation of plaque components for these large population studies could support research with fast and reliable quantitative results saving costs and reducing effort.

There exist some other studies that researched automatic plaque component segmentation. Several studies have used histology data to train supervised classifiers [49][50][51]. However manual annotated MRI data is easier to obtain and have had succes in segmenting plaque components [52][39][53] [54]. All these papers used conventional machine learning algorithm such as Bayes based classifiers, support vector machines, linear and quadratic discriminant classifiers and random forest. Random forest has been a very popular in medical imaging with applications such as: [55][56][57] and was in the study of van Engelen et al. [39] the best performing classifier together with the linear and quadratic discriminant classifiers. Lately deep learning architectures have become increasingly popular for classification tasks and segmentation taks. Recently Dong et al. [58] published a study segmenting plaque components in multi-modal MR Images. In addition to fibrous tissue, lipid rich necrotic core, calcification and haemorrhage they added a fifth class: loose matrix. The authors compare three different deep learning algorithms (GoogLeNet [23], VGG-16 [25] and ResNet-101 [59]) in a study containing over 1000 patients from 13 medical centers all over China. In their comparison they found ResNet-101 outperforming other methods and achieving state-of-the-art results. U-net like ResNet uses skip connections to create deeper networks. Aside from the deep learning methods a conventional method MAPPS [60] was included in their comparison. The study for this method is mentioned as unreliable by van 't Klooster [54] for including training data in the test set. It must be noted that also in Dong et al.[58] only a 80/20 split for training/testing is mentioned. Information about the optimization procedure, loss function and if a validation set is used are lacking. The studies of Dong et al [58] and van Engelen et al. [49] will be used to compare the results to. Even though this is an unfair comparison (different sequences, training set size etc.) it can give us some insight in how this work relates to other scientific work.

3.3 Multiple Instance Learning

Multiple instance learning (MIL) can reduce the need for costly annotations in segmentation tasks by weakening the required degree of supervision. In the multiple instance framework a label per vessel, image or per patient can be used to train classifiers where in the supervised framework a label per pixel would have been required. Dietterich et al. [61] introduced the multiple instance framework with the key chain example. For clarity an adaption of the key-chain example by Babenko [62] is used here:

There are several faculty members, and each owns a key chain that contains a few keys. You know that some of these faculty members are able to enter a certain room, and some are not. The task is then to predict whether a certain key or a certain key chain can get you into this room. To solve this we need to find the key that all the 'positive' key chains have in common. Note that if we can correctly identify this key, we can also correctly classify an entire key chain - either it contains the required key, or it does not.

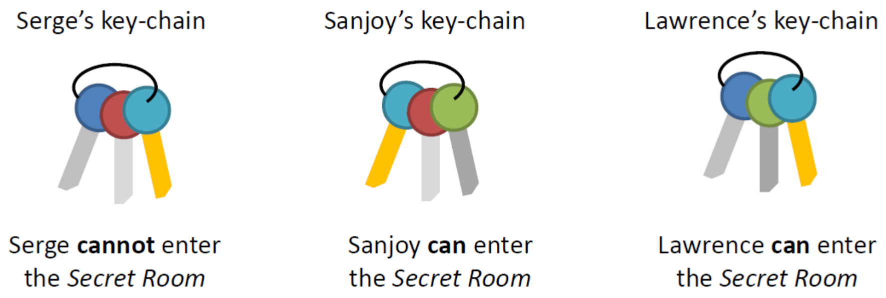

Serge's key-chain
Sanjoy's key-chain
Lawrence's key-chain
Serge **cannot** enter the *Secret Room*
Sanjoy **can** enter the *Secret Room*
Lawrence **can** enter the *Secret Room*

Figure 3.5: The adapted keychain example From Babenko et al. [62]. Only the labels for the key-chains are known. Nonetheless it is still possible to infer which key opens the secret room. Sanjoy's and Lawrence's key-chain can both open the secret room and both have the blue and green key. Serge's key-chain cannot while it has the blue key. Thus we can conclude that the green key opens the secret room.

By comparing the different keys of the key-chains that can open and cannot open the door we can infer which key opens the door (figure 3.5). In the example only 'weak' information was given. Information about the correct label of each key-chain was available but the correct label for the individual keys could only be inferred.

In the multiple instance framework the key-chains are called bags and the keys are called instances. The multiple instance learning is a broad framework, the objective can be to classify the instances as in the example (instance-level classifiers) or to infer the bag labels (bag-level classifiers). Hence it is no surprise that many problems can be written as multiple instance problems or that multiple instance problems can be solved with other frameworks. For example semi-supervised learning is closely related with multiple instance learning. A multiple instance problem can be written as a special case of a semi-supervised problem if negative bags are seen as labeled data (every instance has the label negative) and positive bags are seen as unlabeled data with the constraint that at least one instance is positive [63].

It is clear that in the key-chain example a key-chain is positive if one or more of the keys are positive. As in the example you are able to open the door with one of the keys on a positive key-chain and with none of the keys on a negative key-chain. The translation from instances to bag label is an assumption that differs per problem. The assumption used in the key-chain example is the most common assumption, the standard assumption: a bag is labeled positive if and only if at least one of the instances is positive. Some problems may require other assumptions. Such as the problem of emphysema detection [64] where the following assumption was used: a bag is positive if and only if at least a certain fraction of the instances is positive.

The multiple instance learning framework has been applied in many fields from document-categorization [65], music information retrieval [66], stock selection [67], to many medical applications [68] and it has even beed applied for pornographic image recognition [69]. Some of these applications are inherently a multiple instance problem but part of its popularity is because the multiple instance learning framework allows classifiers to learn with weakly labelled data. In this thesis the focus will be on image level labels. Each image/slice will only be labelled for presence of the components. Learning with weak labels is interesting since it could reduce labelling effort. The internet is a great source of weakly labelled data but also patient records contain vast amounts of weakly labelled data that could be used in the future for training algorithms.

The implementations of all the conventional multiple instance algorithms used in this thesis originate from the Multiple Instance Learning toolbox [70] and are used in combination with the PRTools toolbox classifiers [38] in Matlab.

MIL for Plaque Component Detection

In this thesis the bags are defined as all the pixels within the vessel wall and the lumen in each slice. The bags are thus created per vessel and the instances are thus the pixels within the vessel (figure 3.6). For conventional algorithms only the relevant pixels are given as an input. For the deep multiple instance learning (MIL U-net) a cropped image around the centreline is given as an input. The deep multiple instance learning network is thus supplied with irrelevant pixels from outside the vessel. Therefore the vessel segmentation is added as an additional input to make it easier for the algorithm to learn which pixels are relevant. For the described multiple instance problem vessel segmentation is thus still a prerequisite to create the bags. In this thesis the focus will be on predicting the correct bag labels. The bag labels are thus the presence of plaque components in the vessel in the current slice (examples in figure 3.6).

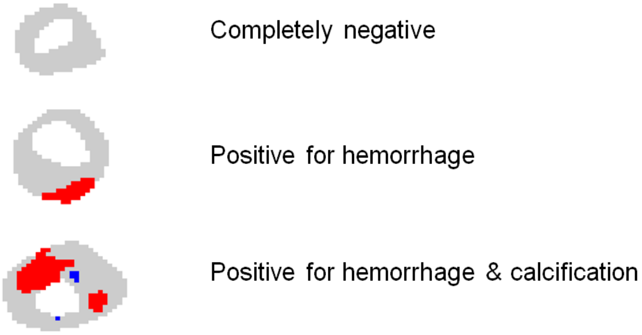


Figure 3.6: Pixel-wise ground truth annotations which visualize the bags and the perfect instance level classification. On the right the corresponding bag labels

More formally, the dataset consists of n bags: $\{B_1, ..., B_n\}$. The i-th bag, B_i contains n_i amount of instances $\{x_{i1}, ..., x_{in_i}\}$. The number of instances (n_i) vary per bag since the instances are the pixels between the lumen and the vessel wall and it's features. The labels of the bags are denoted as y_i while the label of the j-th instance is denoted as y_{ij}. The labels are converted from the pixel-wise annotations using the standard multiple instance assumption (described formally in the next section, equation 3.2). The eventual dataset D is thus defined as: $D = \{(B_i, y_i), ..., (B_n, y_n)\}$ with bags in the form $B_i = \{(x_{i1}, y_{i1}), ..., (x_{in_i}, y_{in_i})\}$.

3.3.1 Simple and Specializing MIL wrappers

Regular algorithms for the supervised learning framework can still be used in the multiple instance framework. The simplest solution, hence called simple MIL, is to use the bag label for every instance. In the negative bags this is correct when the standard multiple instance assumption is used. In the positive bags this introduces false positives, not every instance in a positive bag is positive. By assigning the bag-labels to the instances a dataset is obtained where every example has a label and this data set can thus be trained with a regular fully supervised classifier.

A slightly more sofisticated procedure is the specializing MIL. This is a generalization of the mi-svm [71] procedure. This algorithm initializes with the simple approach. All the instances in positive bags are labelled positive and a classifier is trained. The trained classifier is then used to relabel the instances in the positive bags. While relabelling the data the positive bags should still fulfil the constraint to remain positive i.e. a positive bag should still contain at least one positive instance. If this constraint is violated the label of the least negative example will be changed to positive. With the newly obtained labels the classifier is trained again. This procedure is iterated until the labels converge and stop changing.

The multiple instance assumption will differ per classifier. For the simple and specialized MIL the standard assumption will be used: a bag is positive if and only if at least one instance is positive. A bag label can thus be made with the following combination rule:

$$\forall i : \max_j y_{ij} = y_i \tag{3.2}$$

In words: For every instance in the bag; the label of the most positive instance is the bag label.

3.3.2 MIL-Boost

MIL-Boost by Babenko et al. [72] is a generalization of MIL-Boost by Viola et al. In boosting the goal is to train a strong classifier by combining iteratively multiple weak learners. Weak learners are classifiers whose performance are often only slightly better than guessing. The strong classifier is created by linearly combining the weak learners with weights relative to the weak learners' performance. Training proceeds sequentially and each time a weak learner is trained and added to the strong classifier, the examples are reweighed. Examples that remain incorrectly classified by the strong classifier increase in importance. New weak learners will therefore focus on correctly classifying the previous incorrectly classified examples. In MIL-Boost [72] the weights of the bags is dependent on the weight of its instances and these are adjusted after each iteration. The loss function is defined in the paper as:

$$\mathcal{L}(f) = -\sum_{i=1}^{n_i} \omega_i \left(1(y_i = 1)\log(p_i) + 1(y_i = -1)\log(1 - p_i)\right) \tag{3.3}$$

where f is a weak leaner, y_i is the bag label and ω is the weight for each bag. p_i is the maximum positive posterior probability of all the instances in the bag. The max function is approximated by the differentiable softmax($\cdot$) function.

$$p_i = \text{softmax}_j(p_{ij}) = \text{softmax}_j(\sigma(2f(x_{ij})) \tag{3.4}$$

Where $\sigma(\cdot)$ is the sigmoid function. In this thesis the Noisy-OR rule was used (shown in equation 3.5) instead of the softmax function. The multiple instance assumption is determined by the Noisy-OR rule. Since the Noisy-OR rules allows for more positives before a bag is positives the resulting multiple instance assumption becomes less strict. More information on Noisy-OR can be found in section 3.5.

3.3.3 MIForest

Originally MIForest [73] would be included as a second baseline for the conventional multiple instance algorithms. This would allow to compare a regular random forest with its related multiple instance variant. MIForest has a similar working principle as random forest in the specialized MIL wrapper. Both initialize by labelling the instance labels with the bag labels. In an iterative manner the labels are relabelled until the labels converge (wrapper) or until the heat parameter has cooled down (MIForest). Both use the standard assumption and correct for consistency between bag-labels and instance prediction each iteration by assigning the most positive instance to the positive class. The algorithms differ in the way instances are relabelled. MIForest makes use of deterministic annealing and is inherently multi class. The wrapper is used in a binary classification setting with one class versus the rest approach.

3.3.4 MIL U-net

First related work in combining multiple instance learning with deep learning will be discussed. Inspired by these studies a U-net architecture in the multiple instance framework is proposed.

Related Work

There have been several attempts to integrate convoltional neural networks in the multiple instance learning framework. The most obvious way to combine multiple instance learning with convolutional neural networks is to map the instance-level predictions to bag-level predictions. The standard multiple instance learning framework maps the instance-level predictions with the $\max(\cdot)$ function to bag-level predictions. By implementing such a global max pooling layer the error is only back-propagated through instances considered to be the most likely cause for the bag label by the CNN. This would result in a very high punishment of false positives since one false positive would change the bag label while a false negative is barely punished. This leads to networks that would only find one or a couple of positive instances in a positive bag. Although the standard multiple instance assumption does allign with the diagnostic procedure this would probably not lead to the intended results since in practice radiologist do not look at single pixels. Other multiple instance learning assumptions and thus aggregating functions are probably more fit to mimic the readiologist behaviour for the task in this thesis. Kraus et al. [74] selects other possible functions such as: Noisy-Or [75], integrated segmentation and recognition (ISR) model [76], generalized gean and Log-Sum-Exp (LSE) [77]. Wang et al [69] and Pinheiro et al. [78] used the softmax function [79] and the Log-Sum-Exp aggregation function instead. These functions do approximate the max function and are differentiable speeding up training. Kraus et al [74] tested Noisy-or and ISR and found that both were too sensitive for outliers (false positives) for their application. Therefore the authors defined the Noisy-AND pooling function which models thresholds on instance proportions. The model was extended with a fully connected layer after this MIL pooling layer, to learn the relationship between instances of different classes. For example a relationship that lipid is often found in the same slice as haemorrhage. The MIL pooling layer only learns the relationships between instances of the same class.

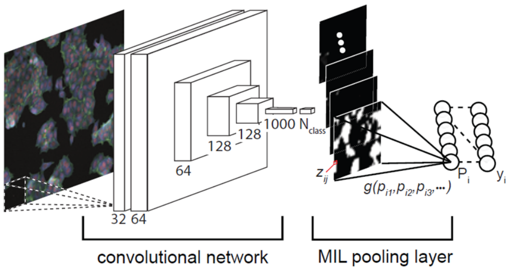


Figure 3.7: The multiple instance learning model from Kraus et al. [74]. A regular convolutional network adapted to the multiple instance framework with a MIL pooling layer. The global pooling function $g(\cdot)$ is aggregating all the instance probabilities p_{ij}. Image retrieved from Kraus et al.[74]

MIL U-net for Carotid Plaque Detection

The MIL U-net is heavily inspired by Kraus et al [74]. The convolutional network in their work is replaced by the U-net architecture. The architecture consists thus of the U-net architecture by Ronneberger et al. [41] followed by the MIL-pooling layer by Kraus et al [74] (figure 3.8). The MIL-pooling layer is added together with a fully connected layer to the U-net architecture. With an identical U-net architecture as in the supervised setting the weights of the layers in the U-net are interchangeable. This allows us to train the supervised U-net with a small data set in a supervised setting, save the network weights, load the weights in the MIL U-net and continue training with weakly labelled data in the multiple instance framework. It is also possible to train first with weakly labelled data in the multiple instance framework and load the U-net layers from the MIL U-net in the fully supervised U-net and continue training in the supervised framework. In this thesis it would make sense to do this with the 20 training patients that have been manually reviewed for errors.

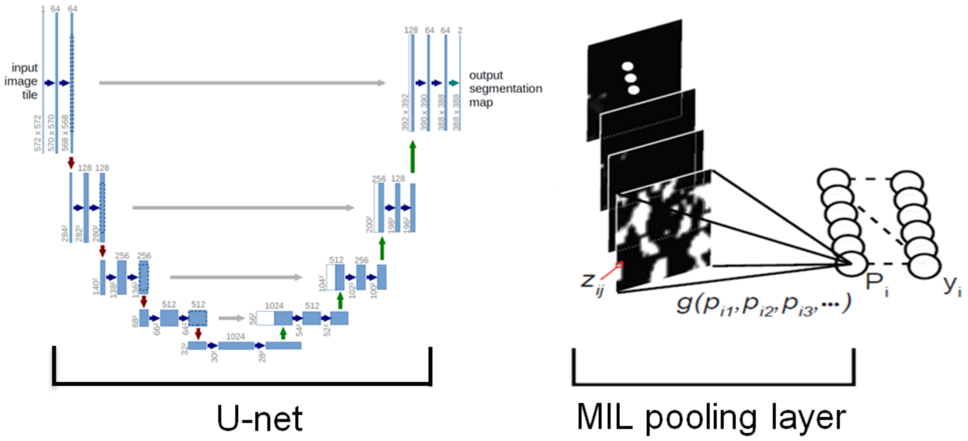


Figure 3.8: MIL U-net, a combination of U-net (Ronneberger et al. [41]) and MIL pooling layers (Kraus et al. [74])

In the MIL pooling layer an aggregating function needs to be chosen. The aggregating function determines how a positive bag is defined by this classifier. The first candidate would be the normal *max()* function. The *max()* function would let this algorithm follow the standard multiple instance assumption: a bag is positive if and only if one or more instances are positive. The standard multiple instance assumption is clinically sound but very strict since one pixel could determine the label of the bag. In practice the radiologist will not base its decision on one pixel. A slightly less stricter candidate is the Noisy-Or:

$$NoisyOr = 1 - \prod_{j}(1 - p_{ij}) \tag{3.5}$$

The third candidate is the NoisyAnd [74] MIL aggregating function:

$$NoisyAnd = \frac{\sigma(a(p_{\bar{ij}} - b_c) - \sigma(-ab_c))}{\sigma(a(1 - b_c) - \sigma(-ab_c)} \tag{3.6}$$

Where $p_{\bar{ij}} = \frac{1}{|j|}\sum_j p_{ij}$. The parameters a and b are learn-able parameters in this thesis. Parameter a controls the slope of the function and parameter b translates the function over the ratio of the feature maps that need to be activated (figure 3.9). Kraus et al [74] fixed the slope of the function and only parameter b was learned. In this thesis a will be a learn-able parameter. NoisyAnd is unique because it can learn the rule the expert used: how many pixels (instances) need to be positive before a slice (bag) is positive. It might give insight in behaviour of the annotator.

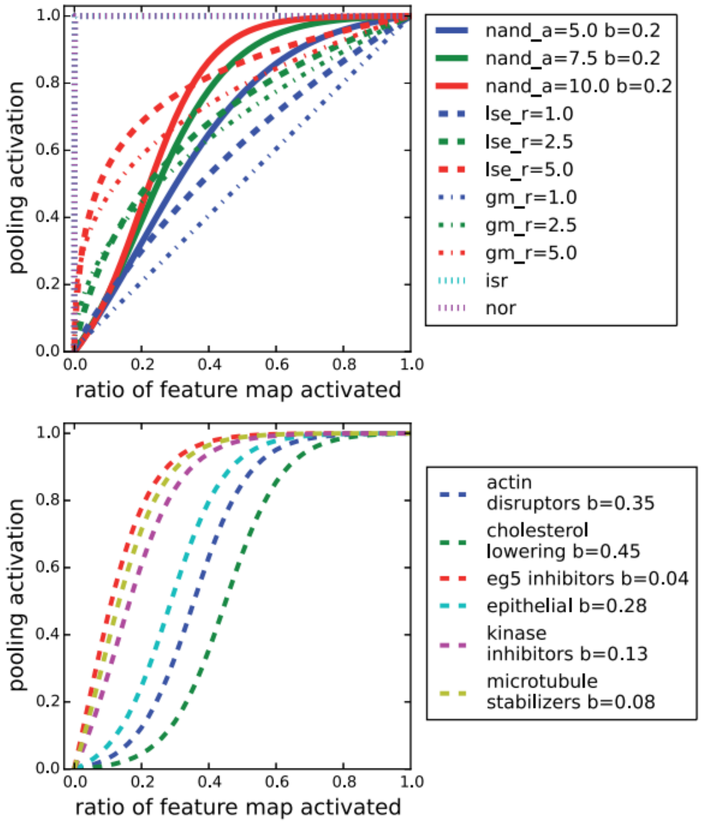


Figure 3.9: Overview of the MIL Pooling functions by Kraus et al. [74]. The top graph shows MIL pooling function candidates and their behaviour, the effect of parameter a on the NoisyAnd function can also be seen. The bottom graph shows the effect of the parameter b on the NoisyAnd function

Obviously there are more candidates for the aggregating function for the MIL-pooling layer as discussed in the related work section 3.3.4 and in figure 3.9 but due to time constraints only the $max()$, Noisy-Or and Noisy-and were tested for the MIL U-net.

With the MIL-pooling layer pooling over all the pixels (instances) in the slice (bag), the prediction for each slice will be a probability per class. In this case the categorical cross-entropy will not suffice, since the problem is a multi-label problem: each image has as many labels as there are classes and multiple classes can co-exist (table 3.2). To cope with this and the class imbalance the loss function has been changed to a weighted binary cross-entropy function.

$$\mathcal{L} = -\sum_i t_i \log(p_i) w_1 + (1 - t_i) \log(1 - p_i) w_2 \tag{3.7}$$

where t_i is a tensor containing the ground truth bag label for each class per slice. With p_i the predicted probability (ranging from 0 to 1) of the bag label per class. Instead of a single weight matrix w the weight matrix is split in two different weight matrices to have more control over the weight and to allow for an asymmetrical weight for false positives and false negatives. To bound the predictions in range zero to one the sigmoid function is used as an activation function.

	F	**L**	**C**	**H**
Bag (Slice) Labels	1	0	0	0
	1	0	0	0
	..	..	..	..
	1	0	1	0
	1	0	0	1
	..	..	..	..
	1	0	0	0
	1	0	0	0
	..	..	..	..
	1	0	0	0
	1	0	0	0

Table 3.2: Example of the labels as used in the loss function. Each row can now contain multiple ones; each image can contain multiple plaque components. The predictions for each component are obtained by pooling over the columns of table 3.1 with the MIL-pooling layer. The predicted probabilities per image together with the labels per image are used to calculate the loss as in equation 3.7.

3.4 Evaluating Performance

Evaluating the performance of the classifiers is not a trivial problem [59]. Especially in the medical field where the data sets are often imbalanced. Diseased tissue, the target class, is often a minority class compared to healthy tissue, the background class. In all experiments in this thesis the background class is reduced by only evaluating the pixels in the region between the vessel wall and lumen. However the background class (fibrous tissue) still contains a vast majority of the pixels. In such an imbalanced data classifying every example to the majority class can result in an overall accuracy over 90% [80]. In this study, classifying everything as fibrous tissue leads to an overall accuracy of roughly 93%. For this reason metrics are calculated over all the classes individually and other metrics are used to evaluate the performance of the algorithms on the imbalanced data set. All the metrics are derived from the confusion matrix.

The metrics used in this thesis are:

- **Accuracy** : the proportion true positives (TP) and true negatives (TN) to the total number of pixels for that class. It is the percentage of correctly classified pixels.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.8}$$

- **Sensitivity**: also called true positive rate, recall or probability of detection. It is a measure to indicate the proportion of positives that are correctly identified as positives. Thus for example the proportion of haemorrhage that is classified as haemorrhage.

$$Sen = \frac{TP}{TP + FN} \tag{3.9}$$

- **Specificity**: also called the true negative rate is the proportion of negatives that are correctly classified as negatives.

$$Spe = \frac{TN}{TN + FP} \tag{3.10}$$

- **Precision**: the proportion of true positives compared to the true positives and false positives.

$$Pre = \frac{TP}{TP + FP} \tag{3.11}$$

- **F_1-score**: the F_1-score is the harmonic mean of the precision and sensitivity. It is also known as the Sørensen-Dice index or the Dice Similarity Coefficient a common metric for evaluating segmentation results.

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{3.12}$$

- **Area under Curve (AUC)**: more correct the area under the receiver operating characteristic curve (ROC curve). In the ROC curve the true and false positive rate are plotted for all possible decision thresholds. The ROC does show the results for different decision thresholds and allows thus for additional optimization for this threshold. Thus the area under the curve (AUC) combines the sensitivity and specificity in a single metric. Random predictions would result in an area under the curve of 0.5 and the line $x = y$. AUC can be used to filter out models that are representative but not discriminative. Note: in this thesis for every class the ROC curve is calculated with an one versus the rest approach which may result in inaccuracies for the AUC and ROC.

- **Intraclass correlation coefficient (ICC) [81]** (A-1): The degree of absolute agreement among measurements made on randomly selected objects. It estimates the correlation of any two measurements.

  Observations are interpreted with the random effects model where observations are seen as a summation of μ an unobserved overall mean, α an unobserved random effect shared by all values in group, and ϵ an unobserved noise term. The ICC is calculated with the variances from these values.

$$ICC = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2} \tag{3.13}$$

- **Cohen's Kappa**: The Intraclass correlation coefficient is a generalization of Cohen's kappa. Where ICC is used for measurements in the continuous domain kappa is used in the categorical domain. It estimates the agreement between observers with the consideration of agreement by chance.

  $$\kappa = 1 - \frac{1 - p_o}{1 - p_e} \tag{3.14}$$

  where p_o is the relative observed agreement (accuracy) among observers and where p_e is the hypothetical probability of agreement by chance. Cohen's kappa is often interpreted based on arbitrary chosen ranges, the most popular being: no agreement $\kappa \leq 0$, slight agreement 0-0.20, fair 0.21-0.40, moderate 0.41-0.60, substantial 0.61-0.80,and (almost) perfect agreement 0.81-1.00

During optimization it is not possible to improve all the metrics simultaneously. Therefore the focus will be to improve the F_1-score, a metric commonly used for evaluating imbalanced data sets [59] [82] and used in similar related work [58] and suggested in [35].

Chapter 4

Experimental Results

4.1 General Experiments

To understand the data set better and to gain better results in the main experiments some preliminary experiments were necessary. One concern could be the split in training, validation and test set. The manual annotated training set, without the extra 93 patients with the automatic vessel segmentations, was chosen to consist of only 20 patients to have a reasonably sized test set. To confirm the size of the training set a learning curve has been made (see section 4.1.2). Similar results as the study from van Engelen et al. [39] are expected, where the learning curve stabilized after 10-15 patients. In the same study the effect of leaving out MRI sequences was studied. The authors concluded that leaving out MRI sequences usually negatively affects results for one or more classes. However leaving out the T2-weighted scans did not have a big impact on their results. Leaving out MRI sequences could significantly reduce pre-processing speed, loading times and storage space. The importance of the MRI sequences is studied in section 4.1.1. Studying the importance of the MRI sequences could also give insight how much the registration problems affect the performance of the pixel-wise classifiers. For pixel-wise classification perfect registration of the sequences is crucial since imperfect registration could assign features from healthy tissue to non-healthy tissue and vice versa, adding noise to the data set.

4.1.1 Sequences

In appendix C an extensive comparison between the different sequences can be found. A random forest classifier was trained on all sequences and for each sequence individually. Since random forest has stochastic elements the random forest was trained three times with the same training and validation set. The mean results on the validation set together with the standard deviations are given in the tables C.1 & C.2. In table C.1 it can be seen that the performance on solely the 3D-FS-Coronal sequence is roughly similar as the performance with all the sequences. The metrics for fibrous and haemorrhage are roughly similar but for the calcification class the classifier trained only on the 3D-FS-Coronal sequence (3D-T1w-gradient echo) clearly outperformed the classifier trained on all the sequences. An explanation can be found in the results in the same table for the other sequences. The classifier does not perform well on any other sequences individually.

In table C.2 one of the sequences is left out during training and classification. The results without 3D-FS-Coronal clearly show that this sequence is of vital importance. Since the training and validation set for both the tables are identical the results between both tables are comparable. In none of the experiments where one of the sequence were left out a better results was obtained as in the experiment with solely the 3D-FS-Coronal sequence. More evidence can be found in appendix B where the feature importance of the experiments with all the sequences are shown. The top 20 features are dominated by features based on the 3D-FS-Coronal (FSCN) sequence only to be contested by the distance based features which are sequence invariant features.

Based on this evidence it was decided to work only with the 3D-FS-Coronal sequence. In all the following experiment the other sequences are discarded. The importance of this choice can not be underestimated. The definition of van Bouwhuijsen et al [21] for the plaque components is clear in the importance of the sequences. Discarding all the other sequences except this sequence means that it is impossible for the classifier to learn the same model as the observers. Haemorrhage and calcification can still be found according to the definition: haemorrhage is defined as hyper-intense in the 3D-FS-Coronal sequence and calcification was defined as a hypo-intense region in all sequence. But the lipid rich necrotic core class is not defined in the 3D-FS-Coronal sequence alone. From earlier work and in related work it became clear that this class is particularly challenging to segment. After careful consideration and after the experts indicated to have more interest in the calcification and haemorrhage class it was decided to fully focus on these two classes and to continue with only the 3D-FS-Coronal sequence. All the following experiments are obtained by solely using the 3D-FS-Coronal sequence.

4.1.2 Amount of Training Data

To evaluate the number of patients necessary in the training set a learning curve has been made which can be found in figure 4.1. In this figure the least stable metrics are shown, sensitivity and F_1-score. Haemorrhage remains quite unstable even after 10-13 patients but shows a more stable prediction from at least 14 patients. The instability can be explained with the fact than not all patients were identified with haemorrhage. In each run the first patient was selected to have all components and every subsequent patient was randomly selected from the training set. The 20 patients chosen in the training set seem borderline sufficient for the haemorrhage class, the other classes stabilize with less patients and show a more stable learning curve. However, this based on the random forest classifier and in the supervised learning framework. More complex classifiers such as deep learning methods may require larger datasets. In multiple instance learning framework while learning with weakly labelled data, a larger training set consisting of more patients is definitely required.

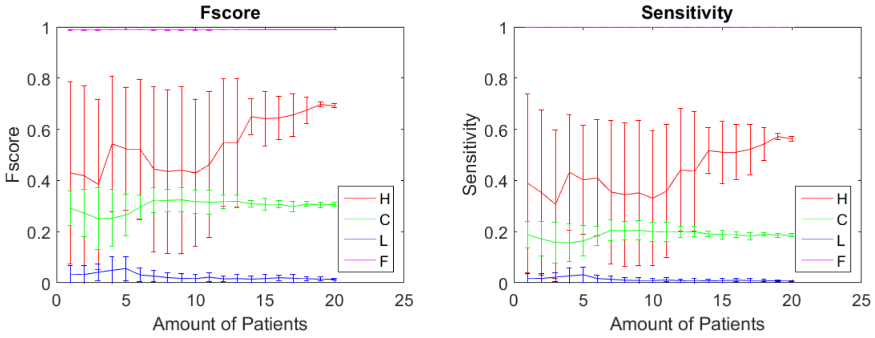


Figure 4.1: Mean values and standard deviations over 5 runs of a learning curve. The first patient was selected on the criteria that every component was present, every subsequent patient was randomly chosen from the training set. Since the training data increases in similarity over the number of patients added, the variance can not be compared between points with a different number of patients in the training set.

4.2 Supervised Classification Results

The supervised results are expected to be the best obtainable results, supervised learning with sufficient data will (nearly) always outperform semi-supervised or unsupervised algorithms. All the classifiers obtained an accuracy of roughly 96%. But for medical diagnostics overall accuracy is not very relevant, since the data set is imbalanced with mostly (healthy) fibrous tissue. Therefore different metrics will be used to measure the performance per class (section 3.4). Table 4.1 and 4.2 show the performance of all the classifiers for the calcification and the haemorrhage class, the two most important classes. A more detailed overview of the performance of the Random forest and the U-net can be found in the following sections 4.2.2 and 4.2.1. A more detailed overview of the results of for the baseline classifiers can be found in Appendix E.

	Comparison Classifiers (20 Patients for Training)				
	Calcification				
	AUC	Sensitivity	ICC	Precision	F_1-score
LDC	0.505	0.468	0.395	0.245	0.322
QDC	0.304	0.508	0.196	0.148	0.202
Loglc	0.536	0.126	0.282	0.537	0.205
RF	0.584 ± 0.002	0.185 ± 0.004	0.388 ± 0.011	0.618 ± 0.013	0.284 ± 0.006
U-net	0.941 ± 0.000	0.623 ± 0.0831	0.226 ± 0.096	0.168 ± 0.041	0.261 ± 0.042
	Haemorrhage				
	AUC	Sensitivity	ICC	Precision	F_1-score
LDC	0.978	0.499	0.729	0.512	0.505
QDC	0.956	0.601	0.261	0.220	0.322
Loglc	0.964	0.426	0.682	0.581	0.491
RF	0.934 ± 0.000	0.420 ± 0.002	0.635 ± 0.006	0.596 ± 0.001	0.493 ± 0.001
U-net	0.917 ± 0.038	0.604 ± 0.071	0.666 ± 0.064	0.422 ± 0.025	0.495 ± 0.014

Table 4.1: Overview of all the fully supervised classifiers for the calcification and haemorrhage class with the mean values and standard deviations over three runs. The baselines linear and quadratic discriminant classifiers (LDC & QDC) and the logistic classifier (Loglc) together with the more extensive optimized random forest (RF) and U-net classifiers.

	Comparison Classifiers (113 Patients for Training)				
	Calcification				
	AUC	Sensitivity	ICC	Precision	F_1-score
LDC	0.292	0.418	0.403	0.251	0.314
QDC	0.228	0.511	0.111	0.103	0.172
Loglc	0.286	0.035	0.071	0.634	0.067
RF	0.517 ± 0.001	0.127 ± 0.006	0.229 ± 0.010	0.667 ± 0.008	0.213 ± 0.009
U-net	0.957 ± 0.010	0.539 ± 0.144	0.627 ± 0.274	0.304 ± 0.081	0.373 ± 0.041
	Haemorrhage				
	AUC	Sensitivity	ICC	Precision	F_1-score
LDC	0.974	0.605	0.641	0.410	0.489
QDC	0.956	0.668	0.177	0.184	0.288
Loglc	0.978	0.393	0.683	0.638	0.487
RF	0.951 ± 0.000	0.467 ± 0.005	0.706 ± 0.006	0.585 ± 0.004	0.519 ± 0.003
U-net	0.977 ± 0.008	0.738 ± 0.036	0.684 ± 0.058	0.394 ± 0.039	0.513 ± 0.028

Table 4.2: Overview of all the fully supervised classifiers for the calcification and haemorrhage class with the mean values and standard deviations over three runs. With the same classifiers as in table 4.1 but with the training set increased by an additional 93 patients. The annotations of these aditional patients are generated from manual annotations of the plaque components by the same observer supplemented with an automatic vessel segmentation obtained by Arias et al. [27].

Table 4.1 and 4.2 show a summary of the results for all the supervised experiments. The classifiers were optimized on the validation set for the F_1-score. The difference in F_1-score between the earlier seen results on the validation set for random forest (appendix C) and the results on the test set (table 4.1 and 4.2) is quite big. This gap in performance between the validation and test set was seen for all of the classifiers. Since the baseline methods did not need tuning it is unlikely that this is a result of over-fitting on the validation set. A new split for training, validation and test set was made and the results for this new split can be found in appendix D. The different scores originate probably form the big variance in the plaque size between patients. This does have a great effect on the ICC score of the volumes per patient, a larger set with more patients, cross validation or a more even distribution with regard of plaque components per patient may give more stable results. Table 4.1 and 4.2 show the mean values with the standard deviation over 3 runs. The baseline classifiers (LDC, QDC and Loglc) are deterministic classifiers where each run returns the exact same results. Random forest and U-net work with random initialization of parameters that cause the algorithm not to converge each run to the exact same results.

The calcification class appears harder to identify for the classifiers than the haemorrhage class. Looking at the calcification class in table 4.1 it can be seen that the classifiers score relatively similar F_1-scores but the means to this score differ for the classifiers. The QDC and U-net do have a relatively high sensitivity and low precision, where the random forest and the logistic classifier have high precision but low sensitivity. The linear discriminant classifier (LDC) has a moderate precision and sensitivity and eventually ends up with the highest F_1-score. The haemorrhage class shows a similar trend with again a relatively high sensitivity and lower precision for the U-net and QDC. The U-net shows promise for the calcification class with a high AUC for calcification especially when comparing to the relatively low area under the curve of the other classifiers. For the haemorrhage class all AUC values are quite good. Although U-net obtains the lowest AUC score, the score is still quite good. Based on the results in table 4.1 LDC seems to be the preferable classifier being the most stable classifier for both classes. U-net shows potential with a high area under the receiver operating characteristic curve.

In the second table 4.2 the results for an incremented training sets are displayed. The training set is increased by 93 patients with annotations created by merging manual annotations of the plaque components of observer one with dilated automatic vessel segmentations obtained with [27]. The additional patients are likely to have registration problems and the automatic vessel segmentation are likely to be of lower quality than human segmentation. Comparing 4.1 to 4.2 show decreased performance for the calcification class for all the conventional methods. This is probably because of the increased amount of noise in the training set and because the distribution of plaque components is different in the additional training patients. Only the deep learning architecture improves consistently by adding additional training data. The U-net is now the best performing classifier for the calcification class. For the haemorrhage class the additional patients in the training set seem to have less drastic effect. But in the end not a lot of improvement can be seen for the conventional methods. Only random forest (a fairly noise resistant algorithm) seem to improve slightly. Likewise the U-net improves slightly, increasing sensitivity while losing precision resulting in a small net gain for the F_1-score.

4.2.1 Random Forest

Random forest has relatively low number of tunable parameters. The most important parameter is the number of trees in the ensemble. This is optimized on the validation set (figure 4.2). The number of features used for splitting the subset in each node is kept the default value of $\sqrt{numberfeatures} = \sqrt{23} \approx 4$ features per node. In all the experiments sequence independent and 3D-FS-coronal features are used as explained in section 4.1.1.

Number of Trees

To select the number of trees necessary for good and stable classification the effect of the number of trees was measured on the validation set. The results of these experiments are plotted in figure 4.2 for the most relevant metrics.

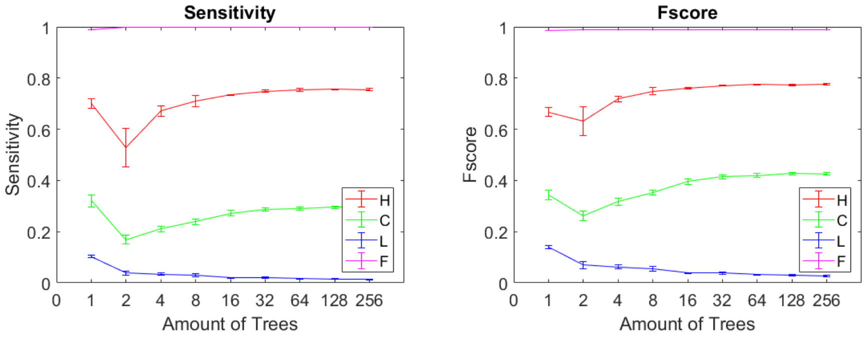


Figure 4.2: Sensitivity and F_1-score as a function of the number of trees. F_1-score is the harmonic mean between sensitivity and precision. Since precision was more stable the fluctuations from sensitivity are damped in the F_1-score.

In figure 4.2 it can be seen that from 16 trees the advantage gained by adding new trees is reduced. To be sure the number of trees chosen is big enough, the eventual number of trees chosen was set to 100 trees. This choice was made with the knowledge that the extra training patients probably introduce more noise to the training data and that having more trees would thus be beneficial for gaining more stable results.

Results Manual Ground Truth

In table 4.3 the results for all the classes obtained with random forest can be found including the validation and test results for fibrous and lipid rich necrotic core class. Random forest selects each node random features and creates a random bootstrap. But with sufficiently large training set and enough trees the algorithm should always converge to roughly the same optimum. This appears to be the case for the runs in table 4.3, inspecting the small standard deviations. The algorithm does not perform well for the lipid rich necrotic core class and standard deviations are higher for this class. Sensitivity for lipid rich necrotic core is the main problem, but sensitivity could not be improved without obtaining a drastically increased number of false positives. For the haemorrhage class there is a quite a gap between the performance in the validation and test set.

Additional experiments with a new split in validation and test set (Appendix D) show that this is probably the result of an uneven split with respect to the plaque-size per patient and the difficulty of the patients.

	Supervised Random Forest (20 Patients for Training)				
	Validation Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.870 ± 0.001	0.999 ± 0.000	0.982 ± 0.000	0.979 ± 0.000	0.989 ± 0.000
L	0.597 ± 0.006	0.007 ± 0.001	0.039 ± 0.005	0.418 ± 0.053	0.014 ± 0.002
C	0.657 ± 0.002	0.184 ± 0.006	0.348 ± 0.006	0.833 ± 0.008	0.302 ± 0.008
H	0.971 ± 0.001	0.564 ± 0.006	0.931 ± 0.005	0.898 ± 0.004	0.693 ± 0.004
	Test Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.893 ± 0.001	0.996 ± 0.000	0.997 ± 0.000	0.981 ± 0.000	0.989 ± 0.000
L	0.583 ± 0.001	0.019 ± 0.004	0.098 ± 0.013	0.175 ± 0.027	0.034 ± 0.007
C	0.584 ± 0.002	0.185 ± 0.004	0.388 ± 0.011	0.618 ± 0.013	0.284 ± 0.006
H	0.934 ± 0.000	0.420 ± 0.002	0.635 ± 0.006	0.596 ± 0.001	0.493 ± 0.001

Table 4.3: Mean and standard deviations over three runs for random forest on the manual annotated data set.

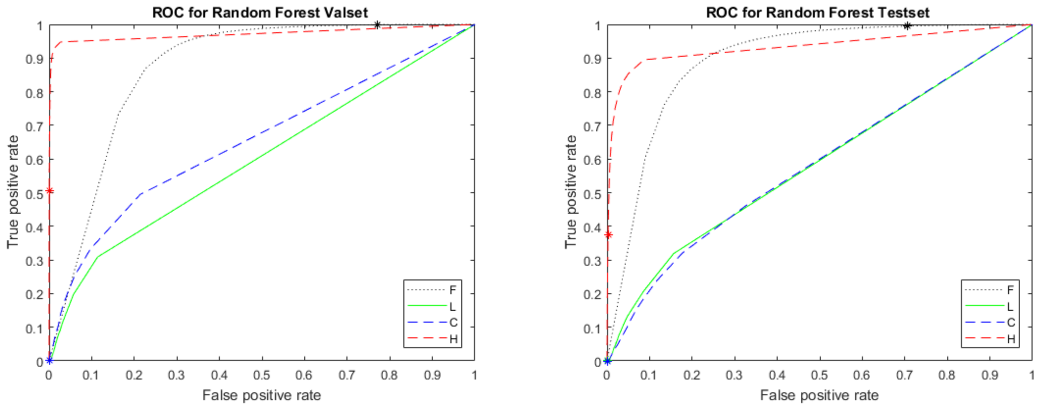


Figure 4.3: ROC curves for the validation and test set of the last run with only the manual annotated patients in the training set. With star the current threshold is denoted. According to this graph the sensitivity (true positive rate) of haemorrhage could be increased with a low cost in false positives by choosing a higher threshold but keep in mind the graph is created with a one-vs-rest approach and contain inaccuracies due to this. From validation to test set the AUC score for calcification decreases quite substantial.

Random Forest: Results Manual & Automatic Vessel Ground Truth

With the addition of extra training patients the performance of haemorrhage seem to improve slightly in the validation and test set. The sensitivity improves slightly together with the ICC score. The ICC would be considered good with a score of 0.706 for the test set. In both sets the F_1-scores for calcification decrease due to a lower sensitivity. The low AUC score for calcification (0.5 is random assignment) shows that random forest is not discriminative for the calcification class.

	Supervised Random Forest (113 Patients for Training)				
	Validation Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.820 ± 0.001	0.999 ± 0.000	0.982 ± 0.000	0.978 ± 0.000	0.989 ± 0.000
L	0.540 ± 0.002	0.000 ± 0.000	0.005 ± 0.002	0.233 ± 0.252	0.001 ± 0.001
C	0.533 ± 0.002	0.139 ± 0.004	0.250 ± 0.009	0.887 ± 0.006	0.240 ± 0.006
H	0.977 ± 0.002	0.605 ± 0.001	0.953 ± 0.002	0.883 ± 0.005	0.718 ± 0.002
	Test Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.884 ± 0.001	0.997 ± 0.000	0.996 ± 0.000	0.981 ± 0.000	0.989 ± 0.000
L	0.560 ± 0.000	0.002 ± 0.001	0.010 ± 0.003	0.172 ± 0.073	0.005 ± 0.002
C	0.517 ± 0.001	0.127 ± 0.006	0.229 ± 0.010	0.667 ± 0.008	0.213 ± 0.009
H	0.951 ± 0.000	0.467 ± 0.005	0.706 ± 0.006	0.585 ± 0.004	0.519 ± 0.003

Table 4.4: Mean and standard deviations over three runs for random forest trained on the manual data set appended with additional patients with automatic segmented vessel annotations

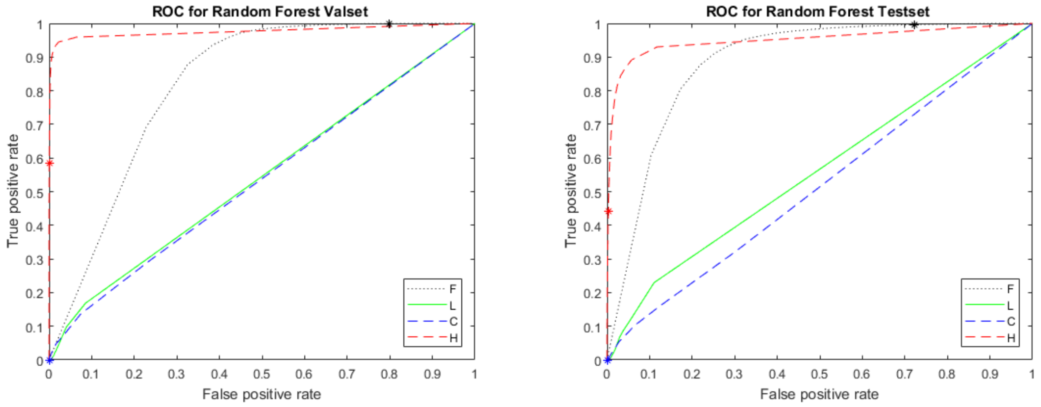


Figure 4.4: ROC curves of the last run with additional patients included in the training set. This includes fully manual annotations as manual component annotations with automatic vessel segmentations. Similar as in the ROC curves for the manual set the threshold for haemorrhage could be optimized with the validation set.

4.2.2 U-Net

The U-net architecture is implemented in the high-level neural network API Keras. The Theano backend was used with GPU implementation using CUDA 8.0 and a NVIDIA GeForce GTX 1070. Training with 20 patients in the training set took approximately 4 hours and training with a training set of 113 patients took approximately 24 hours.

Deep learning architectures are more flexible than most conventional algorithms allowing it to be used in a wide variety of implementations but making it more sensitive to tuning as well. As mentioned in 3.2.3 even the architecture layout is a parameter itself that needs to be tuned. Over 100 experiments were conducted on the validation set to find the most optimal parameters. The tuned parameters are: number of pooling and up-sampling layers (2,3 and 4), number of feature maps (ranged between 16-100), batch size (ranged between 4-64), batch normalization, drop-out and amount of feature maps in the up-sampling path (reduced to 4, the amount of classes).

The optimized layout of the architecture can be found in figure 4.5.

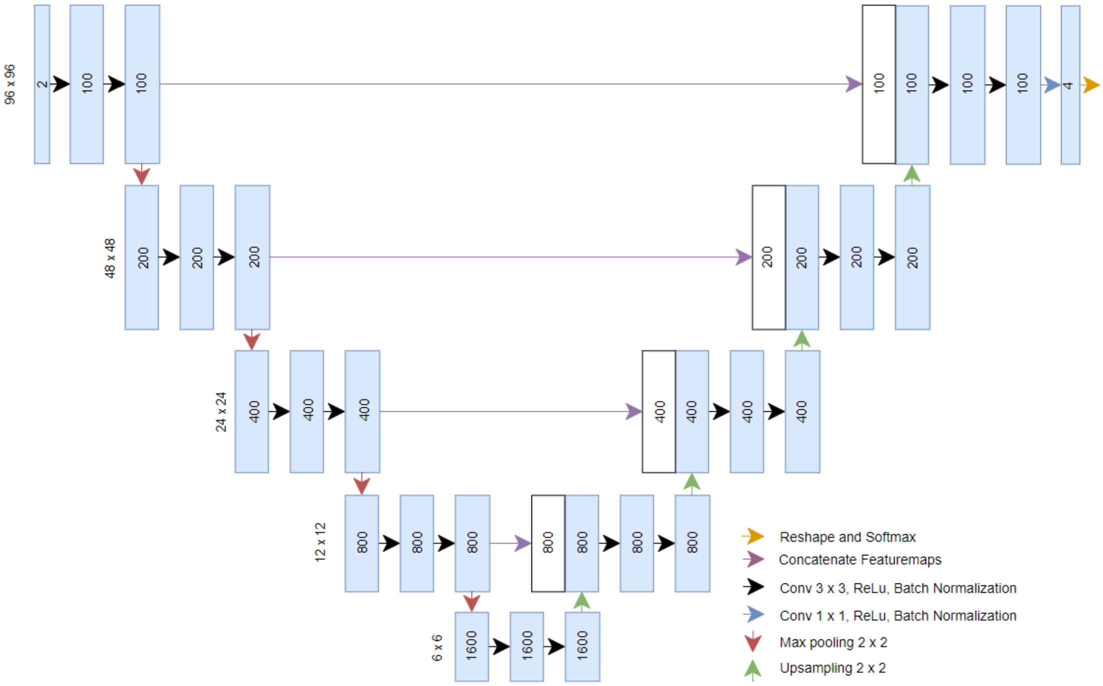


Figure 4.5: The optimized architecture of the U-net. All convolutions are padded, feature maps are displayed in the blocks, while at the left the resolution is displayed.

With 4 pooling and up-sampling layers and batch normalization after each convolution the total number of layers used is 38. Dropout of 50 % was used before pooling and up-sampling. Although the authors from the original batch normalization paper [83] mention that batch normalization could prevent the need for drop-out the results with a combination of both were more favourable. More feature maps appeared to be a more favourable way to spend computational power as opposed to an increased batch size. Batch size was chosen to be 16 and the number of feature maps was set to 100 in the first layers (figure 4.5). The amount of feature maps in the up-sampling layers could be reduced to spare computational power with nearly no cost in performance. The amount of feature maps in the down-sampling layers was of great importance. The size of the network (amount of pooling and up-sampling while continuing the U-net structure) was best for 4 pooling and up-sampling layers but 2 up and down-sampling layers and in combination with more feature maps approached this result. A 'deeper' network over a 'wider' network did thus not result in a big difference but no extreme versions of a deep or wide network were tested.

U-net: Results Manual Ground Truth

Table 4.5 shows the results for all the classes obtained with U-net. Again the gap in performance between the results on the validation set and test set can be seen. U-net has been thoroughly optimized on the validation set and it is therefore unsurprising that the gap in performance is bigger than the drop in performance seen for random forest between the two sets. The intraclass correlation and precision are low for the calcification class. Lipid rich necrotic core is completely ignored and is the only class with a low AUC (figure 4.6). Figure 4.8 shows examples of the segmentation done by the U-net.

	Supervised U-net (20 Patient for Training)				
	Validation Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.765 ± 0.016	0.984 ± 0.006	0.991 ± 0.007	0.984 ± 0.001	0.984 ± 0.002
L	0.537 ± 0.035	0.000 ± 0.000	0.000 ± 0.000	NaN ± NaN	0.000 ± 0.000
C	0.968 ± 0.008	0.618 ± 0.141	0.294 ± 0.143	0.273 ± 0.061	0.370 ± 0.038
H	0.970 ± 0.016	0.814 ± 0.032	0.986 ± 0.007	0.723 ± 0.014	0.766 ± 0.008
	Test Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.830 ± 0.018	0.967 ± 0.011	0.991 ± 0.007	0.987 ± 0.001	0.977 ± 0.005
L	0.627 ± 0.023	0.000 ± 0.000	0.001 ± 0.003	NaN ± NaN	0.000 ± 0.000
C	0.941 ± 0.000	0.623 ± 0.083	0.226 ± 0.096	0.168 ± 0.041	0.261 ± 0.042
H	0.917 ± 0.038	0.604 ± 0.071	0.666 ± 0.064	0.422 ± 0.025	0.495 ± 0.014

Table 4.5: Mean and standard deviations over three runs for U-net trained on 20 completly manual annotated patients

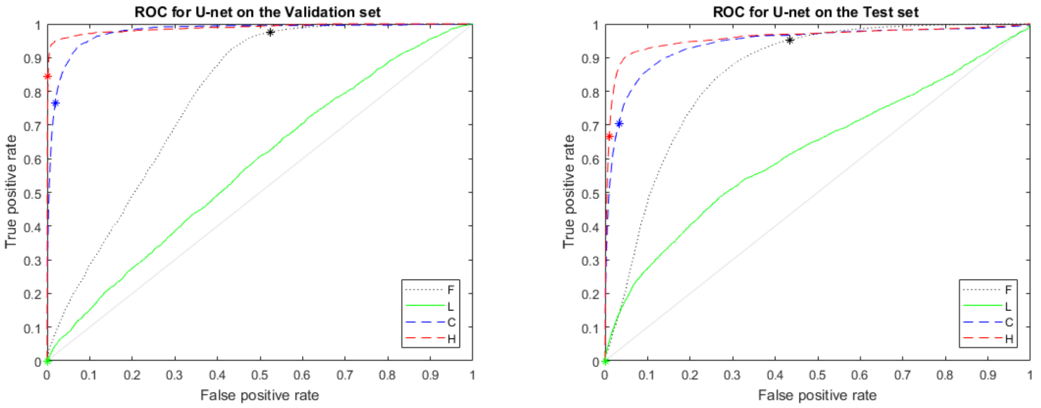


Figure 4.6: ROC curves of the first run with only 20 manual annotated patients in the training set. The operating points can be moved over the ROC curve by choosing another decision threshold. An operating point in the utmost left corner would result in a perfect classification

U-net: Results Manual & Automatic Vessel Ground Truth

In general the results obtained with the incremented training sets are favourable compared to training with only 20 manually annotated patients. The calcification class improves substantially. ICC, sensitivity, precision and as a result F_1-score all improve. Especially the ICC score and precision benefit greatly. Although the ICC score still has room for improvement, the agreement would still be described as good. For the haemorrhage class the precision decreases but sensitivity increases more resulting in a net gain for the F_1-score. The AUC values are quite good and show that U-net can be improved by tuning the decision threshold (figure 4.7). However, the shown AUC values and ROC curves are made with a one-vs-all approach and may therefore be inaccurate.

	Supervised U-net (113 Patients for Training)				
	Validation Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.865 ± 0.011	0.987 ± 0.006	0.995 ± 0.001	0.985 ± 0.001	0.986 ± 0.002
L	0.755 ± 0.019	0.002 ± 0.004	0.005 ± 0.004	NaN ± NaN	0.004 ± 0.008
C	0.983 ± 0.004	0.681 ± 0.113	0.671 ± 0.224	0.397 ± 0.103	0.487 ± 0.062
H	0.994 ± 0.003	0.902 ± 0.011	0.952 ± 0.022	0.624 ± 0.074	0.735 ± 0.049
	Test Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.900 ± 0.008	0.979 ± 0.006	0.998 ± 0.001	0.988 ± 0.001	0.983 ± 0.003
L	0.763 ± 0.016	0.006 ± 0.009	0.016 ± 0.028	NaN ± NaN	0.012 ± 0.016
C	0.957 ± 0.010	0.539 ± 0.144	0.627 ± 0.274	0.304 ± 0.081	0.373 ± 0.041
H	0.977 ± 0.008	0.738 ± 0.036	0.684 ± 0.058	0.394 ± 0.039	0.513 ± 0.028

Table 4.6: Mean and standard deviations over three runs for random forest on the manual data set appended with additional patients in training set

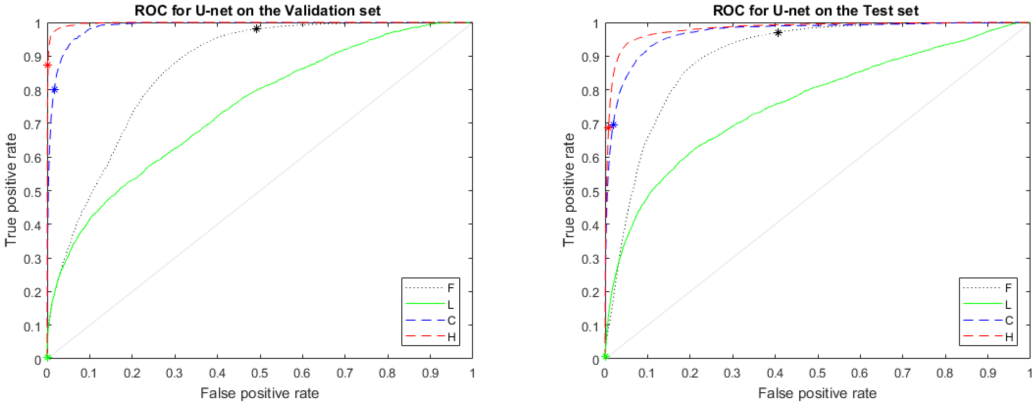


Figure 4.7: ROC curves of the first run with only the manual annotated patients in the training set. Small improvements could be made by choosing a slightly higher threshold for calcification and haemorrhage which would lead to near optimal AUC for the validation set.

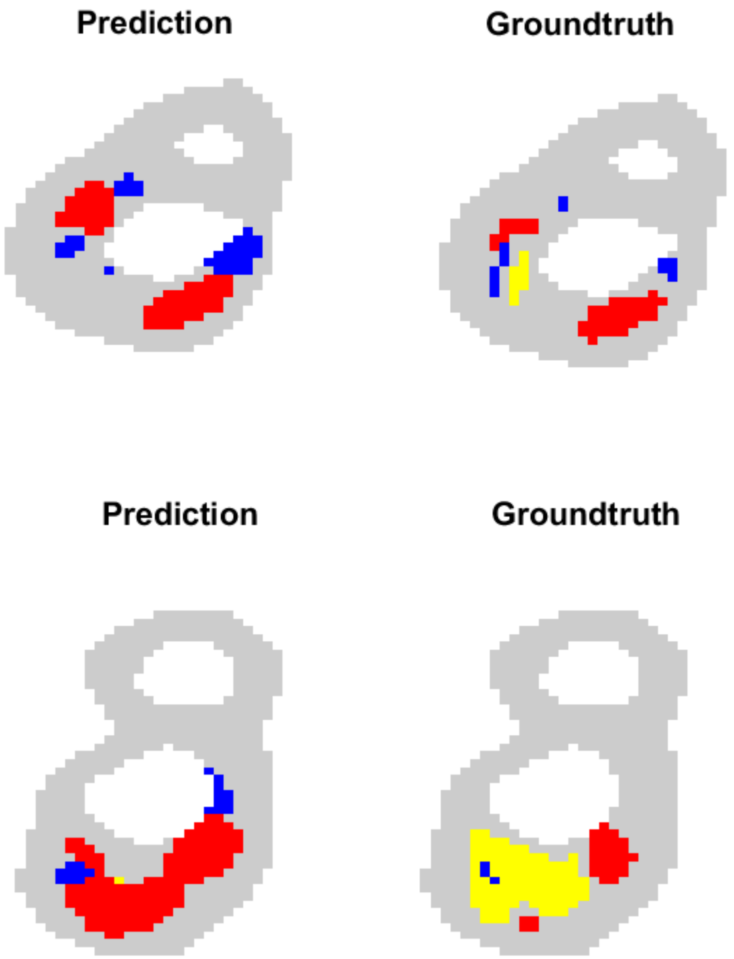


Figure 4.8: Examples of random slices where all components were present. On the left a predictions of the U-net and on the right the accompanying ground truth. Haemorrhage is red, calcification blue and lipid rich necrotic core yellow. In the top row it can be seen that the predictions are close to the ground truth but that the U-net over-segments the calcification and the haemorrhage class. In the bottom row it can be seen that lipid rich necrotic core is classified as haemorrhage with the exception of one single pixel

4.3 Multiple Instance Results

Fibrous tissue class is always present in each slice. All the negative bags should only contain fibrous tissue and positive bags contain at least one other component. Considering the difficulty for the supervised classifiers and the way lipid rich necrotic core is defined the choice was made to focus purely on the calcification and haemorrhage class in the multiple instance learning framework. The calcification and haemorrhage class are classified in a one versus the rest approach for all classifiers except MIL U-net. The training set is the same set as the combined set used in tables 4.2,4.4 and 4.6, consisting of 113 patients. Validation and test set are the same through the whole thesis, consisting of 10 and 21 patients respectively.

To compare the U-net in the supervised framework with the results obtained by the MIL U-net in the multiple instance framework the pixel-wise predictions from table 4.6 are translated to image-level predictions. This is done with the standard multiple instance learning assumption (i.e one or more pixels in a slice predicted as a lipid, calcification or haemorrhage means the bag label is positive for that class). These results for haemorrhage and calcification can be found in table 4.7 in the row U-net alongside the MIL methods.

4.3.1 Conventional MIL methods

The methods used in the MIL wrapper are used with the default parameters of the supervised variant and with the standard multiple instance learning assumption (max-rule). All the conventional methods (LDC, QDC, LOGLC and RF) are trained with the simple and the special MIL wrapper

4.3.2 MIL U-Net

The multiple instance U-net consists of two parts: The U-net architecture and the MIL-pooling layers. The U-net architecture is described 3.2.3 and has been implemented with the parameters found in section 4.2.2. The second part consist of the MIL-pooling layer followed by a fully connected layer.

Multiple candidates for the pooling function are evaluated for the MIL layer, namely the max, Noisy-Or and Noisy-And function (described in section 3.3.4). The Noisy-And was the least sensitive to the weights on the false positives and negatives. By choosing a weight of 100 for calcification and haemorrhage class the first non-trivial results were obtained with the Noisy-And. This led to the choice for the Noisy-And layer and the weights were tuned further for this layer resulting in the following weight matrices:

$$w_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 50 \end{bmatrix} \qquad w_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.1}$$

in the regular class order: fibrous tissue, lipid rich necrotic core, calcification and haemorrhage. w_1 gives weight to false negatives in the loss function and w_2 gives weight to false positives in the loss function. Generally, higher values for w_1 will result in a higher sensitivity while higher weights in w_2 will result in a higher specificity. The effect of the weight-matrices can be deduced from the binary cross entropy loss function:

$$\mathcal{L} = -\sum_i t_i \log(p_i) w_1 + (1 - t_i) \log(1 - p_i) w_2 \tag{4.2}$$

The MIL-pooling layer has 5 interpretable weights learned during training (see equation 3.6). One weight a which controls the slope and parameter b_c is meant to represent an adaptable soft threshold for each class. Where Kraus et al. [74] fixed a and tested with values ranging from 5 to 10, it is a learn-able weight in the MIL U-net. The learned weights for a converged to 1.43, 1.45 and 0.93 in the three runs. The effect of parameter a on the strictness of the multiple instance assumption can be seen in figure 4.9.

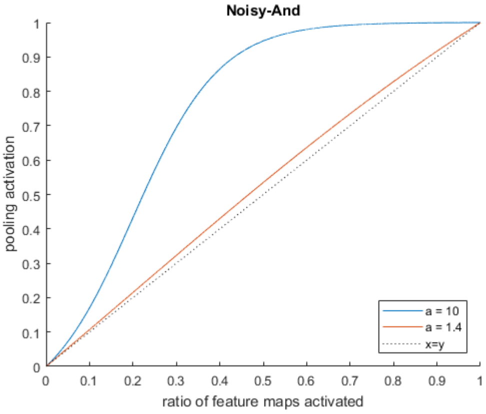


Figure 4.9: The effect of the learn-able parameter a, the higher it is the more 'strict' the assumption becomes. The parameter b was chosen as a constant value of 0.2 (rough mean of the obtained b over all the classes). The obtained slope is closely related with the function $x = y$ simply returning the ratio of feature maps activated. A class is considered positive if its positive probability is higher than 0.5

	Comparison Multiple Instance Classifiers				
	Calcification				
	AUC	Sensitivity	Cohen's kappa	Precision	F_1-score
simple-LDC	0.808	0.706	0.347	0.377	0.491
simple-QDC	0.742	0.872	0.154	0.241	0.377
simple-Loglc	0.856	0.334	0.419	0.825	0.476
simple-RF	0.856 ± 0.003	0.361 ± 0.007	0.445 ± 0.006	0.823 ± 0.004	0.502 ± 0.006
spec-LDC	0.574	0.199	0.235	0.602	0.299
spec-QDC	0.600	0.899	0.114	0.220	0.354
spec-Loglc	0.707	0.074	0.084	0.449	0.128
spec-RF	0.828 ± 0.002	0.783 ± 0.008	0.379 ± 0.007	0.388 ± 0.004	0.519 ± 0.005
MIL-Boost	0.860	0.291	0.391	0.915	0.441
U-net (113 tr.) [2]	0.865 ± 0.005	0.753 ± 0.128	0.400 ± 0.109	0.428 ± 0.109	0.531 ± 0.065
MIL U-net	0.6150 ± 0.051	0.225 ± 0.099	0.180 ± 0.065	0.389 ± 0.035	0.277 ± 0.086
U-net (20 tr.) [1,2]	0.818	0.810	0.283	0.318	0.457
pre-MIL U-net [1]	0.654	0.355	0.240	0.375	0.365
	Haemorrhage				
	AUC	Sensitivity	Cohen's kappa	Precision	F_1-score
simple-LDC	0.935	0.805	0.581	0.514	0.627
simple-QDC	0.870	0.925	0.189	0.187	0.311
simple-Loglc	0.940	0.692	0.717	0.797	0.741
simple-RF	0.942 ± 0.001	0.700 ± 0.013	0.666 ± 0.010	0.693 ± 0.007	0.697 ± 0.010
spec-LDC	0.845	0.780	0.500	0.435	0.559
spec-QDC	0.716	0.956	0.078	0.128	0.226
spec-Loglc	0.878	0.642	0.596	0.626	0.634
spec-RF	0.919 ± 0.002	0.820 ± 0.004	0.502 ± 0.012	0.427 ± 0.012	0.562 ± 0.010
MIL-Boost	0.915	0.522	0.622	0.856	0.648
U-net (113 tr.) [2]	0.970 ± 0.005	0.830 ± 0.023	0.726 ± 0.049	0.695 ± 0.086	0.754 ± 0.043
MIL U-net	0.880 ± 0.020	0.717 ± 0.011	0.399 ± 0.049	0.353 ± 0.043	0.472 ± 0.040
U-net (20 tr.) [1,2]	0.918	0.610	0.687	0.851	0.711
pre-MIL U-net [1]	0.855	0.730	0.440	0.388	0.507

Table 4.7: Metrics for the prediction of the slice labels. Mean and standard deviations are added for algorithms with random elements except for algorithms marked with [1] which denotes the algorithm has been run once. [2] Denotes supervised pixel-wise results obtained with strong labels that are translated with the standard assumption to bag labels. The set-size in patients is added for these classifiers. The weights of U-net (20 tr.) have been used to initialize (a pre-trained) pre-MIL U-net.

The predictions vary a lot between classifiers and are generally better for the haemorrhage class than the calcification class similar as in the fully supervised setting. LDC and QDC in the MIL wrappers do generally have a higher sensitivity and lower precision while the MIL-Boost and the logistic classifier do have a lower sensitivity but higher precision. The random forest in the simple and specialized wrapper do display different behaviour but obtain the highest F_1-scores for the calcification class. Simple-RF by conservative behaviour with moderate sensitivity and high precision and random forest in the specialized wrapper with high sensitivity and moderate precision. The simple-RF and simple-Loglc even have a higher agreement with the experts (Cohen's kappa) than the supervised U-net (113 training patients) trained with strong labels. For the haemorrhage class the logistic linear classifier in the simple MIL wrapper obtains the best performance. The performance of instance-level predictions for the haemorrhage and calcification class by the linear logistic classifier in the simple wrapper can be found in appendix F.

The proposed MIL U-net is outperformed by most methods. Even when the weights are loaded from a supervised U-net, trained with 20 pixel-wise labelled patients, the performance is lower than most classifiers. By loading a pre-trained layer the performance increased but with the same layers a supervised U-net with a global pooling layer would have been more effective (manual U-net in the table). An additional multiplication with the vessel segmentation with the feature maps needs to be added to ensure only the pixel within the vessel are weighted. In addition the parameter for the strictness of the assumption (figure 4.9) need to be evaluated to see if it got stuck in local optima or if it is not learn-able at all.

Chapter 5

Discussion

In this thesis supervised and multiple instance classifiers have been applied to segment plaque components in the carotid artery. Based on the comparison between the supervised classifiers no specific classifier is found to be the best for all components. Overall all the components U-net seems to be the most suitable classifier after the addition of extra patients in the training set. It is the only classifier with a good ICC inter-rater agreement of 0.627 for calcification. Good classification results were obtained for the haemorrhage class with high AUC for all classifiers. U-net obtained the second best F_1-score of 0.513 and has a good ICC inter-rater agreement of 0.684 for the haemorrhage class.

The performance of the multiple instance classifiers was also better for the haemorrhage than for the calcification class. The logistic classifier in the simple wrapper outperforms all the other classifiers trained with weak labels for this class. All the simple MIL classifiers with the exception of quadratic Bayes perform quite well for the haemorrhage class. For the calcification class all the supervised classifiers perform better in the simple wrapper than in the specialized MIL wrapper with the exception of random forest. It would therefore be interesting to see how the closely related MIForest [73] would perform since this has a similar working principle but it has been designed to cope with multiple instance problems. Although it must be noted that a regular supervised logistic classifier trained with one-sided noise in the simple MIL wrapper beats classifiers designed to cope with a multiple instance problem such as MIL-Boost and MIL U-net. These supervised classifiers trained with weak labels approach the performance of U-net trained with strong labels, for the haemorrhage class the difference in F_1-score between the simple-Loglc and the U-net is less than 2%.

5.1 Supervised Learning

The patients were carefully divided over the sets based on the relative amount of components in the sets. Nonetheless there was quite a big gap in the performance between the validation and test set for the supervised results. For classifiers optimized thoroughly on the validation set this is expected, as parameters of the classifiers have been chosen to fit the validation set the best.

However the default parameters were used for LDC, QDC and Loglc. All these classifiers show a decreased performance in the test set compared to the to the validation set. It appears that the test set is harder to classify than the validation set. Visual inspection shows that scan quality and difficulty between patients can differ greatly. More patients may need to be included to gain a more balanced data set and a more stable overview of the performance of the classifiers.

Sequences

One of the most important decisions was to continue with only the 3D-FS-Coronal sequence. This decision is based on the experiments in appendix B and C. These experiments show overwhelming evidence for the importance of the 3D-FS-Coronal sequence and show that the other sequences do not add extra information. The results of these experiments are in conflict with the conclusion of van Engelen et al. [39] that leaving out different MRI sequences usually negatively affects results for one or more classes. The other sequences in this study are not as infomative as the sequences used in the study of Van Engelen [39] (which included pre- and post contrast scans) or registration problems do negate the extra information available in these sequences. This is an important difference of this study compared to related studies which nearly always include multiple sequences. It is suspected that the observers preferred the 3D-FS-Coronal sequence and used this sequence to annotate most of the data. This could lead to a bias for the 3D-FS-Coronal sequence since there would be less registrations errors for this sequence in the annotations.

Lipid Rich Necrotic Core

Lipid rich necrotic core is not defined with solely the 3D-FS-Coronal sequence. However, even when all the sequences were included the classifiers were unable to make adequate predictions for the lipid class. An explanation can be found in the observed agreement for the lipid class between observers. The absolute agreement of 57.67% and 71.57% for the patient-wise presence would probably decrease further for pixel-wise agreement. More informative sequences [84], contrast agent [85] [86] or new features are necessary for adequate lipid rich necrotic core predictions. The substantial higher agreement between observers for the other classes indicates that the lipid rich necrotic class is the hardest class to identify.

The best predictions for lipid rich necrotic core are made by the LDC (appendix E.1) with a sensitivity of 17 %, an ICC score of 0.48 (a moderate agreement with the observer) and a F_1-score of 0.136. The study by Engelen et al.[39] agrees that lipid rich necrotic core is the hardest class to identify. In this study the maximum sensitivity for this class was 13% with an maximum ICC score of 0.88. However, the maximum sensitivity for the lipid rich necrotic core class was 6% on the Erasmus Medical Center data. The accompagnying maximum ICC value of 0.41 is similar to the ICC found in this study. The postcontrast sequence was contributed most information for the lipid class in this study. In van Engelen and al. [39] and in this thesis the performane of LDC was considerably better for the lipid class than random forest. The state-of-the-art results are found by

Dong et al. [58], reaching a F_1-score of 0.520. In this study the T2W sequence contributed most information.

Calcification

Calcification was better identified than the lipid rich necrotic core. There is quite a difference in the behaviour between the tested classifiers for this class. With the logistic classifier and random forest with more conservative predictions resulting in low sensitivity and higher precision. The linear and quadratic discriminant classifiers which are less conservative resulting in a higher sensitivity but a lower precision due to a higher number of false positives. U-net has the highest sensitivity but a lower precision. This algorithm has been tuned with weights for false negatives while optimizing for the F_1-score. To increase the precision for this algorithm an extra term could be included in the loss function placing more weight on the false positives. Results for calcification could improve by obtaining better registrations or new annotations for 3D-FS-coronal sequence. For pixel-wise classification perfect registration is required and for the calcification this is even more crucial since its structure is often small and narrow. As a consequence registration errors will introduce relatively more noise compared to other classes. This can be seen if patients with automatic vessel ground truth are added to the training set (this set will contain more registration errors). The performance of all the classifiers decrease with the exception of U-net.

The best performing classifiers for the calcification class is the U-net with a relatively high sensitivity of 53.9 % and a good intraclass correlation coefficient. But the F_1-score shows that in precision improvements should be made. Compared to Dong et al. [58] a lower F_1-score is obtained (0.373 to 0.580). Compared to van Engelen at al. [39] the sensitivity of the classifier is higher 53.9% to the maximum in this study of 35 %. The associated ICC score of 0.627 for the U-net is lower than the ICC score of 0.84 for the random forest found in their study.

Haemorrhage

Haemorrhage is the best performing class with in general higher F_1-scores and ICC values. The performance slightly decreases with the addition of the extra patients for most classifiers with the exception of random forest and U-net. Random forest and U-net both perform well with the additional patients in the training set. Random forest with a lower sensitivity (46.7% to 73.8 %) but higher precision (58.5% to 39.4%) and higher intraclass correlation coefficient (0.706 to 0.684) seems the best classifier for haemorrhage.

Compared to van Engelen et al. [39] with a near perfect ICC scores of 0.97 the ICC scores are lower but the maximum sensitivity of 0.63 is exceeded by the U-net. The F_1-score is again lower (0.519 for random forest, 0.513 for U-net) compared to the 0.671 obtained by the ResNet-101 in Dong et al.[58]. However, the GoogLeNet and VGG-16 in the same paper score only 0.580 with their data set of 1098 patients fully annotated patients. 20% out of the 1098 patients were used as test set while using the remaining for training (with no mention of a validation set). In this paper the authors found that one sequence (MP-RAGE) was dominating the haemorrhage classification, achieving higher F_1-score on this sequence alone. In this thesis the performance of the haemorrhage was also dominated by one sequence (3D-FS-Coronal) as discussed in section 4.1.1.

5.1.1 Deep learning versus Conventional Learning

In tables 4.1 and 4.2 the results can be compared for the calcification and the haemorrhage class between U-net and the conventional machine learning algorithms. Only U-net improves for all classes with the additional patients added to the training set. Random forest improves for the haemorrhage class but has a decreased performance for the calcification class. Deep learning methods are notorious for requiring huge amounts of data and with the addition of 93 patients to the training set the U-net seems to have the best overall results. The benefit of extra training examples outweighed the extra noise these less refined examples introduce. Another good explanation for the drop in performance for the conventional classifiers could be found in the feature importance (appendix B). The second and third most important features are distance based features. The automatic vessel segmentations are likely less accurate than the manual annotations and are dilated to include the component annotations and to prevent under-segmentation of the vessel. This could be the cause of the decreased performance for the conventional classifiers when the 93 patients with automatic vessel segmentations were added to the training set. The distance based features became less accurate for predicting plaque components. The importance of the distance based features could also be used to improve the U-net. For every slice fed to the U-net a (binary) mask is supplied in the second channel with the segmentation of the vessel. The purpose of this mask is to let the U-net focus on the pixels within the vessel wall and lumen. Considering the strength of the distance based features a logical improvement could be to replace the binary vessel segmentation mask with a mask with information about the distance to the nearest vessel wall. In many other deep learning applications the mask is directly applied to the image to blacken everything outside the region of interest. This is not done in this study since the vessel is relatively small and border effects could have a great (negative) impact. In general deep learning methods will outperform conventional machine learning algorithm given enough data. Features created with domain-expertise could be given to deep learning algorithms as well to improve performance further.

5.2 Multiple Instance Learning

For both classes the random forest and the logistic classifier trained with one-sided noise (Simple MIL) perform quite well. For the random forest this can be explained by the bias random forest has for the majority class [82]. This is favorable when random forest is applied in the multiple instance setting wherein false positives have more influence than false negatives. Examples of supervised classifiers trained with one-sided noise outperforming multiple instance algorithms can be found in literature as well. For example Ray et al. [87] and Alpaydin et al[88] concluded that whether a multiple instance learner outperforms its supervised counterpart in the simple-MIL depends on the domain of the data set and the type of learner. This is in line with the conclusion of Corbonneau et al.[89] who concluded that for instance classification tasks where the witness rate (the proportion of positive instances in positive bags) is relatively high the problem can be cast as a regular supervised problem with one-sided noise. In the experiments of van Winckelen et al. [90] fully supervised algorithms trained in a naive way (Simple MIL wrapper: propagating the bag labels), frequently approached the multiple instance methods on multiple instance problems and frequently outperformed them. The authors also concluded that better performance on bag level does not necessarily imply that instance level performance is better and the other way around. This should also be considered for the translation of the supervised results of U-net to bag level predictions. The results for the supervised U-net are optimized on instance level without regard for the bag-level predictions. Since the multiple instance learning algorithms use different assumptions the instance-level performance is not included in this thesis. The true value of multiple instance learning is hard to evaluate without the instance-level performance in this application. Multiple instance learning achieves reasonable performance overall on bag-level and some supervised classifiers trained in the multiple instance framework approached and outperformed the fully supervised trained U-net in bag-level predictions. Multiple instance learning is thus a viable option for screening, obtaining image/slice based predictions with more easily obtainable weakly labelled data. Extensive study on instance-level performance is necessary to determine the value of multiple instance learning and weakly labelled learning for the pixel-wise segmentation task.

5.2.1 U-net versus MIL U-net

Unfortunatly MIL U-net has not been optimized thoroughly. Inspecting figure 4.9 it becomes clear that the Noisy-And layer learns a very loose multiple instance assumption. Kraus et al.[74] experimented with three different values $a = 5, 7.5$ and 10 for the steepness of the slope, which reflects the strictness of the assumption. In this thesis the parameter a was made a learn-able weight in the layer initialized with $a = 1$. It is possible that the network is unable to learn a good value for the slope by getting stuck in local optima. Experiments where the slope is initialized with higher values or is fixed as in [74] are necesarry to determine if the network is able to learn a fitting weight for the slope of the function. To test this efficiently the imported layers from U-net could be frozen to learn only the optimal parameters of the MIL pooling layers. The NoisyAnd function in the MIL-pooling layer could give a unique insight, the observer's rule for labelling could be deduced. It must also be noted that Kraus et al. [74] formulated a joint cross entropy function with terms for both the MIL-pooling layer and the output of the network. (fully connected layer). The inclusion of the MIL-pooling layer in the loss function could help optimizing the MIL-pooling layer. Another improvement necessary is the multiplication of the vessel segmentation input with the feature maps before the MIL pooling layer. This will ensure that only the pixels within the vessel are weighted. After training the MIL-pooling layers could be removed or the weights from the U-net part of the network could be transferred to a U-net. The obtained network can be used to obtain pixel-wise (instance-level) predictions. These pixel-wise predictions are necessary to evaluate the benefit of deep multiple instance learning for the segmentation task. MIL U-net is unique for being able to train in supervised and a multiple instance setting but when weights are loaded from a trained supervised network the eventual predictions should be at least better (on bag level) than the predictions from the loaded supervised network.

5.3 Limitations

Registration errors and incorrect annotations as a consequence of registration errors introduced noise to a degree it was better to exclude the least informative sequences. Since the annotations are based on all the sequences extra noise is introduced by the annotations itself as well. While the manual ground truth was visually reviewed to filter out registration errors, the annotations added with the automatic vessel ground truth were not due to the size and time limitations. In attempt to filter out registration errors maximum translation during rigid registration was not allowed to be bigger than 2 mm. This seems to filter out some registration errors but the set remained noisy. The vessel segmentations used for the automatic vessel ground truth are based on the PDw-FSE-BB (black blood), T2w-EPI and the phase contrast MRI. Hence, it is expected that the 93 extra training patients with automatic vessel segmentations introduce extra noise as a result of registration errors. All the sequences are required for identifying plaque components from a clinical point of view since some components are not (well) defined in a single sequence. Especially the lipid rich necrotic core class is not defined in the used 3D-FS-Coronal sequence.

The data set is highly imbalanced with respect to the fibrous tissue class, increasing the difficulty drastically. Previous work during the internship included balancing of the data set but this resulted in an increased number of false positives and lower overall accuracy. Weighting each class differently had a beneficial effect for the U-net but the similar weights did not improve the results for the random forest. Optimizing for each classifier the weights for each class would be too time-consuming. The weights for each class in the U-net can still improve when more time is spend fine-tuning the weights. Based on the ROC curves most algorithms can be improved as well by changing the decision threshold. This is not done since the used ROC are based on a one versus the rest classification and due to time limitations. Multi-class problems in general are harder to optimize than binary classification task since optimizing for one class can result in a decreased performance of other classes.

5.4 Future work

It would be interesting to do a more extensive study in the value of each sequence in the Rotterdam Study for the different plaque components. If lipid rich necrotic core could be visualized and defined in the 3D-FS-Coronal sequence, pre-processing could be simplified and registration problems could be circumvented. Furthermore, pixel-wise annotations by multiple obsevers needs to be obtained to evaluate the performance with repect to human observers and to compare inter-observability. The training set could still be extended with additional patients. Deep learning methods will benefit from this and other deep learning methods could still be explored such as SegNet [91], DilatedNet [92], Deeplab v3 [93] and of course the 3D U-net [94]. The 3rd dimension could help getting more consistent classification over the slices. The H-DenseUNet [95] hybrid densely connected U-net is a hybrid between a 2D and a 3D network. A 2D DenseUNet is used for efficiently extracting features within the slices and is combined with a 3D DenseUnet which is responsible for the extracting hierarchical features from the information in the slices and the information between slices. Circumventing the high computational cost of 3D convolutions. Finally, post-processing could increase classification performance and increase consistency by removing outliers. Conditional Random Fields [96] could be used to smooth the segmentation based on the image intensities and improve performance further.

Optimizing MIL U-net is necessary and could give insight in how strict the multiple instance assumptions should be. The MIL U-net can improved by adding a multiplication before the MIL-pooling layer to only weight the pixels in the vessels and by initializing the MIL pooling parameters better. Investigating the required MIL-pooling parameters could be done by loading a pre-trained MIL U-net and freeze the U-net layers. Freezing the layers reduces the amount of free parameters, speed up training and make it easier to find suitable MIL-pooling layer parameters. MIL U-net, like U-net, will likely benefit much of increasing the training set size. The data set could be increased up to 585 patient for the current bag choice. Choosing different bag types could be more convenient in practice and could open the opprtunity to use larger (annotated) data sets. Choosing the complete vessel as a bag opens up the annotations of Selwaness et al. [48] consisting of 1663 annotated patients (more than 3000 vessels). With increased bag-size instance-level predictions will become harder. Bag-level predictions could be used for screening and selecting subsets of patients. Bag-wise features could be derived such as volume of the vessel and maximum distance between vessel wall and lumen to improve bag-wise predictions for conventional multiple instance learning methods. More bag-based multiple instance classifiers should be tested with these features if bag-based classification is a priority (screening).

Additionally, it has been shown that with the same U-net architecture vessel segmentation is possible. The segmentation in appendix A was obtained without any optimization. A tuned 3D U-net [94] should easily obtain better results and could possibly outperforming the currently used vessel segmentation tool [27].

Chapter 6

Conclusion

6.1 Conclusion

In this thesis extensive comparisons have been made between fully supervised and multiple instance classifiers. Related work by Arna van Engelen [35], in supervised learning, has been extended to MRI data of the Rotterdam Study and deep learning has been included. With a training set of 20 patients conventional machine learning techniques outperformed U-net, the deep learning architecture. After an augmentation of the training set, 93 patients were added with automatic vessel segmentations provided by Arias et al. [27], the U-net showed the best overall performance. With a sensitivity of 90.2 % and an ICC score of 0.952 in the validation set, a sensitivity of 73.8 % and an ICC of 0.684 in the test set, it is evident that it is possible to segment haemorrhage in the carotid artery. Calcification with a sensitivity of 53.9 % and a ICC score of 0.627 in the test set performs well and these results are comparable with other studies [53][39]. The final class, lipid rich necrotic core, is ill-defined in the chosen sequence and is not comparable to related studies. However this choice for the 3D-FS-Coronal sequence is well supported in various experiments in this thesis. The results of these experiments are in conflict with the conclusion of van Engelen et al. [39] that leaving out different MRI sequences usually negatively affects results for one or more classes. An explanation for this could be found in the differences in sequences between studies and the registration errors encountered between sequences in this study.

In the second part of this thesis an extensive comparison between multiple instance algorithms have been made and a deep multiple instance algorithm, MIL U-net, has been proposed. Multiple instance learning can reduce labelling effort to image level labelling and can potentially open up new easily obtainable data sets in medical imaging in the form of electronic patient records. The labelling has been reduced to vessel-wise labelling per slice and supervised and multiple instance algorithms have been trained to predict these labels. Supervised methods trained with one-sided noise outperformed multiple instance classifiers such as MIL-Boost and the proposed MIL U-net. This is in agreement with studies such as Corbonneau et al. [89] who concluded that for instance classification tasks where the witness rate (the proportion of positive instances in positive bags) is relatively high the problem can be cast as a regular supervised problem with one-sided noise. In addition is shown that it is possible to train the MIL U-net with a combination of supervised and multiple instance learning and that this can improve MIL performance. However the proposed MIL U-net requires more optimization and should, with pre-training in a supervised setting, at least match the bag level performance of the supervised network used to pre-train.

Appendix A

Vessel segmentation

Segmentation of the vessel using a similar structured U-net as described in section 3.2.3 with parameters as in section 4.2.2. With only small changes such as ending with one feature map, using a sigmoid instead of a softmax activation and by changing the loss function to binary cross entropy to suit the now binary classification problem:

$$\mathcal{L} = -\sum_i t_i w_1 \log(p_i) + (1 - t_i) w_2 \log(1 - p_i) \tag{A.1}$$

With estimated guesses for the weights (w_1 was set to 100 and w_2 to 1) a Dice score of 0.43 on the test set was obtained. As can be seen in figure A the score could easily be improved by choosing a lower weight for the positive class. This U-net is a feasible candidate for vessel-segmentation. With a 3D U-net [94], a proper split and optimization state-of-the-art results could be achieved.

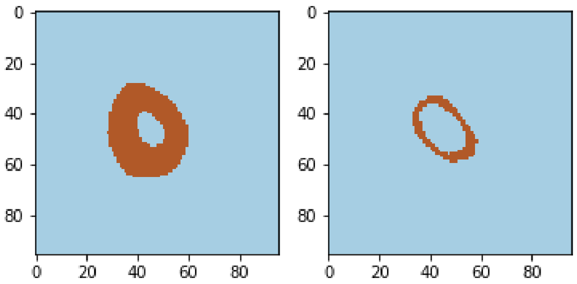


Figure A.1: Example of a segmentation of the vessel by a 2D U-net architecture trained on only 20 patients. Left a random chosen slice with the predicted vessel with right the ground truth. The predicted value is over-segmented most likely due to w_1 being to high.

Appendix B

Feature Importance

Rank	**Featurename**	**Feature Importance**
1	FSCNBlur1	1398.38
2	LumenDistance	1204.34
3	ProductDistances	1103.61
4	FSCNsubimage	1084.77
5	FSCNLapl3	816.29
6	FSCNLapl2	637.56
7	FSCNLapl1	432.91
8	FSCNGM1	405.70
9	FSCNGM2	378.24
10	FSCNGM3	334.73
11	FSCNRfilt	331.56
12	FSCNLocalStd	328.20
13	FSCNGM	317.81
14	FSCNLapl	310.34
15	WallDistance	276.47
16	FSCNGy	230.78
17	FSCNgmag	229.31
18	FV60RNGM3	222.06
19	PDWFSEBBMRANLapl3	205.44
20	PDWEPINLapl3	203.39

Table B.1: The top 20 features ranked by feature importance, for a single run with random forest using all sequences. It is clear that features derived from the 3D-FS-Coronal (FSCN) sequence and the distance based features are the most important features. The first feature based on another sequence, the phase constrast sequence (FV60), is ranked eighteenth. The phase contrast sequence is mainly used by experts for locating the lumen.

Appendix C

MRI Sequence Importance

	Results Random Forest Different Sequences		
	All sequences		
	LRNC	Calcification	Haemorrhage
Accuracy	0.988 ± 0.000	0.993 ± 0.000	0.997 ± 0.000
Sensitivity	0.001 ± 0.001	0.127 ± 0.004	0.560 ± 0.007
Specificity	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.350 ± 0.058	0.864 ± 0.024	0.902 ± 0.004
F_1-score	0.003 ± 0.001	0.221 ± 0.006	0.691 ± 0.005
	FSCN		
	LRNC	Calcification	Haemorrhage
Accuracy	0.988 ± 0.000	0.993 ± 0.000	0.997 ± 0.000
Sensitivity	0.007 ± 0.001	0.184 ± 0.006	0.564 ± 0.006
Specificity	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.418 ± 0.053	0.833 ± 0.008	0.898 ± 0.004
F_1-score	0.014 ± 0.002	0.302 ± 0.008	0.693 ± 0.004
	PDWFSE		
	LRNC	Calcification	Haemorrhage
Accuracy	0.987 ± 0.000	0.992 ± 0.000	0.993 ± 0.000
Sensitivity	0.008 ± 0.000	0.004 ± 0.001	0.001 ± 0.001
Specificity	0.999 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.076 ± 0.007	0.778 ± 0.134	NaN ± NaN
F_1-score	0.015 ± 0.001	0.007 ± 0.001	0.002 ± 0.002
	PRDWEPI		
	LRNC	Calcification	Haemorrhage
Accuracy	0.987 ± 0.000	0.992 ± 0.000	0.993 ± 0.000
Sensitivity	0.007 ± 0.002	0.003 ± 0.000	0.000 ± 0.000
Specificity	0.999 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.084 ± 0.017	0.548 ± 0.041	NaN ± NaN
F_1-score	0.013 ± 0.004	0.005 ± 0.000	0.000 ± 0.000
	T2W		
	LRNC	Calcification	Haemorrhage
Accuracy	0.987 ± 0.000	0.992 ± 0.000	0.993 ± 0.000
Sensitivity	0.006 ± 0.001	0.000 ± 0.000	0.003 ± 0.002
Specificity	0.999 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.113 ± 0.013	NaN ± NaN	0.764 ± 0.105
F_1-score	0.011 ± 0.002	0.000 ± 0.000	0.006 ± 0.004

Table C.1: Results on the validations set with random forest for the different available sequences. Reported values are the mean values over 3 runs with the same train and validation set.

	Results Random Forest without Sequences		
	All sequences		
	LRNC	Calcification	Haemorrhage
Accuracy	0.988 ± 0.000	0.993 ± 0.000	0.997 ± 0.000
Sensitivity	0.001 ± 0.001	0.127 ± 0.004	0.560 ± 0.007
Specificity	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.350 ± 0.058	0.864 ± 0.024	0.902 ± 0.004
F_1-score	0.003 ± 0.001	0.221 ± 0.006	0.691 ± 0.005
	without FSCN		
	LRNC	Calcification	Haemorrhage
Accuracy	0.988 ± 0.000	0.992 ± 0.000	0.993 ± 0.000
Sensitivity	0.002 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
Specificity	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.266 ± 0.051	NaN ± NaN	NaN ± NaN
F_1-score	0.004 ± 0.001	0.000 ± 0.001	0.000 ± 0.000
	without FV60		
	LRNC	Calcification	Haemorrhage
Accuracy	0.988 ± 0.000	0.993 ± 0.000	0.997 ± 0.000
Sensitivity	0.002 ± 0.001	0.137 ± 0.002	0.572 ± 0.005
Specificity	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.431 ± 0.064	0.856 ± 0.009	0.902 ± 0.003
F_1-score	0.003 ± 0.001	0.236 ± 0.002	0.700 ± 0.004
	without PDWFSE		
	LRNC	Calcification	Haemorrhage
Accuracy	0.988 ± 0.000	0.993 ± 0.000	0.997 ± 0.000
Sensitivity	0.002 ± 0.000	0.136 ± 0.006	0.550 ± 0.002
Specificity	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.399 ± 0.089	0.869 ± 0.017	0.906 ± 0.003
F_1-score	0.004 ± 0.001	0.236 ± 0.010	0.685 ± 0.001
	without PDWEPI		
	LRNC	Calcification	Haemorrhage
Accuracy	0.988 ± 0.000	0.993 ± 0.000	0.997 ± 0.000
Sensitivity	0.003 ± 0.001	0.128 ± 0.007	0.564 ± 0.010
Specificity	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.473 ± 0.105	0.876 ± 0.013	0.906 ± 0.007
F_1-score	0.006 ± 0.002	0.223 ± 0.011	0.695 ± 0.008
	without PDWFSE		
	LRNC	Calcification	Haemorrhage
Accuracy	0.988 ± 0.000	0.993 ± 0.000	0.997 ± 0.000
Sensitivity	0.002 ± 0.001	0.139 ± 0.006	0.560 ± 0.005
Specificity	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Precision	0.408 ± 0.121	0.866 ± 0.015	0.902 ± 0.004
F_1-score	0.004 ± 0.002	0.239 ± 0.009	0.691 ± 0.004

Table C.2: Results on the validations set with random forest while leaving out one sequence. Reported values are the mean values over 3 runs with the same train and validation set.

Appendix D

Different Split

	Validation set				
	Calcification				
	AUC	Sensitivity	ICC	Precision	F_1-score
LDC	0.378	0.585	0.441	0.254	0.354
QDC	0.242	0.715	0.131	0.117	0.201
Loglc	0.457	0.158	0.414	0.588	0.249
RF	0.592 ± 0.006	0.254 ± 0.002	0.538 ± 0.010	0.659 ± 0.012	0.367 ± 0.000
U-net	0.938	0.470	0.624	0.297	0.364
	Haemorrhage				
	AUC	Sensitivity	ICC	Precision	F_1-score
LDC	0.978	0.554	0.878	0.510	0.531
QDC	0.953	0.679	0.176	0.161	0.260
Loglc	0.977	0.398	0.841	0.711	0.510
RF	0.951 ± 0.002	0.440 ± 0.006	0.800 ± 0.007	0.646 ± 0.005	0.523 ± 0.006
U-net	0.982	0.865	0.404	0.280	0.424
	Test set				
	Calcification				
	AUC	Sensitivity	ICC	Precision	F_1-score
LDC	0.462	0.523	0.769	0.346	0.416
QDC	0.278	0.584	0.358	0.181	0.277
Loglc	0.495	0.105	0.377	0.593	0.179
RF	0.601 ± 0.005	0.212 ± 0.005	0.402 ± 0.002	0.705 ± 0.017	0.326 ± 0.008
U-net	0.930	0.360	0.924	0.414	0.385
	Haemorrhage				
	AUC	Sensitivity	ICC	Precision	F_1-score
LDC	0.978	0.501	0.933	0.718	0.590
QDC	0.962	0.723	0.898	0.365	0.485
Loglc	0.973	0.382	0.749	0.884	0.533
RF	0.969 ± 0.001	0.368 ± 0.005	0.690 ± 0.005	0.903 ± 0.003	0.523 ± 0.005
U-net	0.975	0.848	0.960	0.586	0.693

Table D.1: Results for the algorithms when a different split between patients for the training, validation and test set is made. U-net and random forest have parameters tuned on patients in the former validation set. These patients are divided over the other sets and the results for these algorithms are thus biased. U-net has been trained and evaluated once.

Appendix E

Results per classifier

E.1 Linear Bayes Normal Classifier

	Supervised Linear Bayes				
	Validation Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.682	0.991	0.998	0.985	0.988
L	0.660	0.169	0.675	0.403	0.239
C	0.501	0.439	0.671	0.395	0.416
H	0.887	0.757	0.995	0.773	0.765
	Test Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.886	0.976	0.995	0.987	0.981
L	0.537	0.173	0.482	0.112	0.136
C	0.505	0.468	0.395	0.245	0.322
H	0.978	0.499	0.729	0.512	0.505

Table E.1: Results for the linear Bayes normal classifier trained with 20 fully manually annotated patients

E.2 Quadratic Bayes Normal Classifier

	Supervised Quadratic Bayes				
	Validation Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.875	0.963	0.965	0.987	0.975
L	0.595	0.268	0.516	0.166	0.205
C	0.368	0.432	0.170	0.202	0.275
H	0.989	0.867	0.849	0.414	0.560
	Test Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.891	0.917	0.900	0.992	0.953
L	0.518	0.335	0.146	0.060	0.101
C	0.304	0.508	0.148	0.126	0.202
H	0.956	0.601	0.261	0.220	0.322

Table E.2: Results for the quadratic Bayes normal classifier trained with 20 fully manually annotated patients

E.3 Logistic Linear Classifier

	Supervised Logistic Classifier				
	Validation Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.905	0.999	0.984	0.978	0.989
L	0.638	0.014	0.210	0.183	0.026
C	0.635	0.121	0.164	0.819	0.210
H	0.984	0.557	0.929	0.893	0.686
	Test Set				
	AUC	Sensitivity	ICC	Precision	F_1-score
F	0.898	0.996	0.997	0.981	0.989
L	0.452	0.037	0.205	0.215	0.063
C	0.536	0.126	0.282	0.537	0.205
H	0.964	0.426	0.682	0.581	0.491

Table E.3: Results for the logistic linear classifier trained with 20 fully manually annotated patients

Appendix F

Simple-Loglc Instance results

	Instance Results Simple-Loglc	
	Calcification	Hemorrhage
Accuracy:	0.993	0.992
Sensitivity:	0.170	0.449
Specificity:	0.999	0.997
Precision:	0.642	0.629
Recall:	0.170	0.449
F_1-score:	0.269	0.524

Table F.1: Instance-level results on the test set for the logistic linear classifier trained with weak labels in a naive way (propagating bag-labels to instance-labels). The training set consisted of 113 patients with bag-level annotations. A bag was defined as the region within the vessel wall and lumen in a slice. The classification approach was one-vs-all which could result in one pixel being classified as both haemorrhage and calcification, but in the multi-class approach these classes where nearly not confused by any classifier. Since the lipid class was not taken in account the metrics may seem better.

Bibliography

[1] World Health Organization. The top 10 causes of death, 2015.

[2] Jacob Fog Bentzon, Fumiyuki Otsuka, Renu Virmani, and Erling Falk. Mechanisms of plaque formation and rupture. *Circulation research*, 114(12):1852–1866, 2014.

[3] John R Guyton and Keith F Klemp. Development of the lipid-rich core in human atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology*, 16(1):4–11, 1996.

[4] JR Guyton and KF Klemp. Transitional features in human atherosclerosis. intimal thickening, cholesterol clefts, and cell loss in human aortic fatty streaks. *The American journal of pathology*, 143(5):1444, 1993.

[5] Frank D Kolodgie, Herman K Gold, Allen P Burke, David R Fowler, Howard S Kruth, Deena K Weber, Andrew Farb, LJ Guerrero, Motoya Hayase, Robert Kutys, et al. Intraplaque hemorrhage and progression of coronary atheroma. *New England Journal of Medicine*, 349(24):2316–2325, 2003.

[6] HR Underhill, C Yuan, VL Yarnykh, B Chu, M Oikawa, L Dong, NL Polissar, GA Garden, SC Cramer, and Thomas S Hatsukami. Predictors of surface disruption with mr imaging in asymptomatic carotid artery stenosis. *American Journal of Neuroradiology*, 31(3):487–493, 2010.

[7] Renu Virmani, Jagat Narula, and Andrew Farb. When neoangiogenesis ricochets. *American heart journal*, 136(6):937–939, 1998.

[8] Renu Virmani, Frank D Kolodgie, Allen P Burke, Aloke V Finn, Herman K Gold, Thomas N Tulenko, Steven P Wrenn, and Jagat Narula. Atherosclerotic plaque progression and vulnerability to rupture. *Arteriosclerosis, thrombosis, and vascular biology*, 25(10):2054–2061, 2005.

[9] Tobias Saam, Holger Hetterich, Verena Hoffmann, Chun Yuan, Martin Dichgans, Holger Poppert, Thomas Koeppel, Ulrich Hoffmann, Maximilian F Reiser, and Fabian Bamberg. Meta-analysis and systematic review of the predictive value of carotid plaque hemorrhage on cerebrovascular events by magnetic resonance imaging. *Journal of the American College of Cardiology*, 62(12):1081–1091, 2013.

[10] Moeen Abedin, Yin Tintut, and Linda L Demer. Vascular calcification. *Arteriosclerosis, thrombosis, and vascular biology*, 24(7):1161–1170, 2004.

[11] Terence M Doherty, Kamlesh Asotra, Lorraine A Fitzpatrick, Jian-Hua Qiao, Douglas J Wilkin, Robert C Detrano, Colin R Dunstan, Prediman K Shah, and Tripathi B Rajavashisth. Calcification in atherosclerosis: bone biology and chronic inflammation at the arterial crossroads. *Proceedings of the National Academy of Sciences*, 100(20):11201–11206, 2003.

[12] Giuseppe Sangiorgi, John A Rumberger, Arlen Severson, William D Edwards, Jean Gregoire, Lorraine A Fitzpatrick, and Robert S Schwartz. Arterial calcification and not lumen stenosis is highly correlated with atherosclerotic plaque burden in humans: a histologic study of 723 coronary artery segments using nondecalcifying methodology. *Journal of the American College of Cardiology*, 31(1):126–133, 1998.

[13] Kiran R Nandalur, Erol Baskurt, Klaus D Hagspiel, C Douglas Phillips, and Christopher M Kramer. Calcified carotid atherosclerotic plaque is associated less with ischemic symptoms than is noncalcified plaque on mdct. *American Journal of Roentgenology*, 184(1):295–298, 2005.

[14] Wael E Shaalan, Hongwei Cheng, Bruce Gewertz, James F McKinsey, Lewis B Schwartz, Daniel Katz, Dindcai Cao, Tina Desai, Seymour Glagov, and Hisham S Bassiouny. Degree of carotid plaque calcification in relation to symptomatic outcome and plaque inflammation. *Journal of vascular surgery*, 40(2):262–269, 2004.

[15] Carl-Magnus Wahlgren, Wei Zheng, Wael Shaalan, Jun Tang, and Hisham S Bassiouny. Human carotid plaque calcification and vulnerability. *Cerebrovascular diseases*, 27(2):193–200, 2009.

[16] Allard C van der Wal and Anton E Becker. Atherosclerotic plaque rupture–pathologic basis of plaque stability and instability. *Cardiovascular research*, 41(2):334–344, 1999.

[17] Aloke V Finn, Masataka Nakano, Jagat Narula, Frank D Kolodgie, and Renu Virmani. Concept of vulnerable/unstable plaque. *Arteriosclerosis, thrombosis, and vascular biology*, 30(7):1282–1292, 2010.

[18] Scott M. Grundy. Metabolic syndrome pandemic. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 28(4):629–636, 2008.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[20] David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[21] Quirijn JA van den Bouwhuijsen, Meike W Vernooij, Albert Hofman, Gabriel P Krestin, Aad van der Lugt, and Jacqueline CM Witteman. Determinants of magnetic resonance imaging detected carotid plaque components: the rotterdam study. *European heart journal*, 33(2):221–229, 2011.

[22] QJA van den Bouwhuijsen, MW Vernooij, BFJ Verhaaren, HA Vrooman, WJ Niessen, GP Krestin, MA Ikram, OH Franco, and A van der Lugt. Carotid plaque morphology and ischemic vascular brain disease on mri. *American Journal of Neuroradiology*, 38(9):1776–1782, 2017.

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[24] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] Albert Hofman, Guy GO Brusselle, Sarwa Darwish Murad, Cornelia M van Duijn, Oscar H Franco, André Goedegebure, M Arfan Ikram, Caroline CW Klaver, Tamar EC Nijsten, Robin P Peeters, et al. The Rotterdam Study: 2016 objectives and design update. *European journal of epidemiology*, 30(8):661–708, 2015.

[27] Andres Arias, Jens Petersen, Arna van Engelen, Hui Tang, Mariana Selwaness, Jacqueline CM Witteman, Aad van der Lugt, Wiro Niessen, and Marleen de Bruijne. Carotid artery wall segmentation by coupled surface graph cuts. In *International MICCAI Workshop on Medical Computer Vision*, pages 38–47. Springer, 2012.

[28] R van 't Klooster, Marius Staring, Stefan Klein, Robert M Kwee, M Eline Kooi, Johan HC Reiber, Boudewijn PF Lelieveldt, and Rob J van der Geest. Automated registration of multispectral mr vessel wall images of the carotid artery. *Medical physics*, 40(12), 2013.

[29] Mariana Selwaness. *Magnetic Resonance Imaging of Carotid Atherosclerosis.* PhD thesis, Erasmus MC: University Medical Center Rotterdam, 2014.

[30] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[31] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2002.

[32] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[33] Alan M Turing. Intelligent machinery, a heretical theory. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, 105, 1948.

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[35] Arna Engelen. *Multimodal Image Analysis for Carotid Artery Plaque Characterization.* PhD thesis, Erasmus MC: University Medical Center Rotterdam, 2014.

[36] Abhishek Jaiantilal. Classification and regression by randomforest-matlab. *URL http://code. google. com/p/randomforest-matlab*, 2009.

[37] Andy Liaw and Matthew Wiener. The randomforest package. *R News*, 2(3):18–22, 2002.

[38] Robert PW Duin, P Juszczak, P Paclik, E Pekalska, D de Ridder, DMJ Tax, and S Verzakov. Prtools 4-a matlab toolbox for pattern recognition. version 4.1. *Delft University of Technology*, 2007.

[39] Arna van Engelen, Marleen de Bruijne, Torben Schneider, Anouk C van Dijk, M Eline Kooi, Jeroen Hendrikse, Aart Nederveen, Wiro J Niessen, and Rene M Botnar. Evaluating classifiers for atherosclerotic plaque component segmentation in mri. In *Annual Conference on Medical Image Understanding and Analysis*, pages 156–168. Springer, 2017.

[40] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[42] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

[43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[44] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[45] Chun Yuan, Lee M Mitsumori, Marina S Ferguson, Nayak L Polissar, Denise Echelard, Geraldo Ortiz, Randy Small, Joseph W Davies, William S Kerwin, and Thomas S Hatsukami. In vivo accuracy of multispectral magnetic resonance imaging for identifying lipid-rich necrotic cores and intraplaque hemorrhage in advanced human carotid plaques. *Circulation*, 104(17):2051–2056, 2001.

[46] Glenn M LaMuraglia, James F Southern, Valentin Fuster, Howard L Kantor, et al. Magnetic resonance images lipid, fibrous, calcified, hemorrhagic, and thrombotic components of human atherosclerosis in vivo. *Circulation*, 94(5):932–938, 1996.

[47] Vincent C Cappendijk, Kitty BJM Cleutjens, Alfons GH Kessels, Sylvia Heeneman, Geert Willem H Schurink, Rob JTJ Welten, Werner H Mess, Mat JAP Daemen, Jos MA van Engelshoven, and M Eline Kooi. Assessment of human atherosclerotic carotid plaque components with multisequence mr imaging: initial experience. *Radiology*, 234(2):487–492, 2005.

[48] Mariana Selwaness, Quirijn van den Bouwhuijsen, Robbert S van Onkelen, Albert Hofman, Oscar H Franco, Aad van der Lugt, Jolanda J Wentzel, and Meike Vernooij. Atherosclerotic plaque in the left carotid artery is more vulnerable than in the right. *Stroke*, 45(11):3226–3230, 2014.

[49] Arna Van Engelen, Wiro J Niessen, Stefan Klein, Harald C Groen, Hence JM Verhagen, Jolanda J Wentzel, Aad van der Lugt, and Marleen de Bruijne. Multi-feature-based plaque characterization in ex vivo mri trained by registration to 3d histology. *Physics in Medicine and Biology*, 57(1):241, 2011.

[50] Sharon E Clarke, Vadim Beletsky, Robert R Hammond, Robert A Hegele, and Brian K Rutt. Validation of automatically classified magnetic resonance images for carotid plaque compositional analysis. *Stroke*, 37(1):93–97, 2006.

[51] JMA Hofman, WJ Branderhorst, HMM Ten Eikelder, VC Cappendijk, S Heeneman, ME Kooi, PAJ Hilbers, and BM ter Haar Romeny. Quantification of atherosclerotic plaque components using in vivo mri and supervised classifiers. *Magnetic resonance in medicine*, 55(4):790–799, 2006.

[52] IM Adame, RJ Van Der Geest, BA Wasserman, MA Mohamed, JHC Reiber, and BPF Lelieveldt. Automatic segmentation and plaque characterization in atherosclerotic carotid artery mr images. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 16(5):227–234, 2004.

[53] Arna van Engelen, Anouk C van Dijk, Martine TB Truijman, Ronald van't Klooster, Annegreet van Opbroek, Aad van der Lugt, Wiro J Niessen, M Eline Kooi, and Marleen de Bruijne. Multi-center mri carotid plaque component segmentation using feature normalization and transfer learning. *IEEE transactions on medical imaging*, 34(6):1294–1305, 2015.

[54] Ronald Van‘t Klooster, Andrew J Patterson, Victoria E Young, Jonathan H Gillard, Johan HC Reiber, and Rob J van der Geest. An objective method to optimize the mr sequence set for plaque classification in carotid vessel wall images using automated image segmentation. *PloS one*, 8(10):e78492, 2013.

[55] Rémi Cuingnet, Raphael Prevost, David Lesage, Laurent D Cohen, Benoît Mory, and Roberto Ardon. Automatic detection and segmentation of kidneys in 3d ct images using random forests. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–74. Springer, 2012.

[56] Ramon Casanova, Santiago Saldana, Emily Y Chew, Ronald P Danis, Craig M Greven, and Walter T Ambrosius. Application of random forests methods to diabetic retinopathy classification analyses. *PLoS One*, 9(6):e98587, 2014.

[57] Soumya Ghose, Jhimli Mitra, Arnau Oliver, R Martí, Xavier Lladó, Jordi Freixenet, Joan C Vilanova, Désiré Sidibé, and Fabrice Meriaudeau. A random forest based classification approach to prostate segmentation in mri. *MICCAI Grand Challenge: Prostate MR Image Segmentation*, 2012, 2012.

[58] Yuxi Dong, Yuchao Pan, Xihai Zhao, Rui Li, Chun Yuan, and Wei Xu. Identifying carotid plaque composition in mri with convolutional neural networks. In *Smart Computing (SMARTCOMP), 2017 IEEE International Conference on*, pages 1–8. IEEE, 2017.

[59] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[60] Fei Liu, Dongxiang Xu, Marina S Ferguson, Baocheng Chu, Tobias Saam, Norihide Takaya, Thomas S Hatsukami, Chun Yuan, and William S Kerwin. Automated in vivo segmentation of carotid plaque mri with morphology-enhanced probability maps. *Magnetic Resonance in Medicine*, 55(3):659–668, 2006.

[61] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1):31–71, 1997.

[62] Boris Babenko. Multiple instance learning: algorithms and applications. *View Article PubMed/NCBI Google Scholar*, 2008.

[63] Zhi-Hua Zhou and Jun-Ming Xu. On the relation between multi-instance learning and semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1167–1174. ACM, 2007.

[64] Isabel Pino Pena, Veronika Cheplygina, Sofia Paschaloudi, Morten Vuust, Jesper Carl, Ulla Moller Weinreich, Lasse Riis Ostergaard, and Marleen de Bruijne. Automatic emphysema detection using weakly labeled hrct lung images.

[65] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, pages 577–584, 2003.

[66] Michael I Mandel and Daniel PW Ellis. Multiple-instance learning for music information retrieval. In *ISMIR*, pages 577–582, 2008.

[67] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576, 1998.

[68] Gwenolé Quellec, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard. Multiple-instance learning for medical image and video analysis.

[69] Yuhui Wang, Xin Jin, and Xiaoyang Tan. Pornographic image recognition by strongly-supervised deep multiple instance learning. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 4418–4422. IEEE, 2016.

[70] Cheplygina V. Tax, D.M.J. MIL, a Matlab toolbox for multiple instance learning, Jun 2016. version 1.2.1.

[71] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 577–584, 2003.

[72] Boris Babenko, Piotr Dollár, Zhuowen Tu, and Serge Belongie. Simultaneous learning and alignment: Multi-instance and multi-pose learning. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[73] Christian Leistner, Amir Saffari, and Horst Bischof. Miforests: Multiple-instance learning with randomized trees. *Computer Vision–ECCV 2010*, pages 29–42, 2010.

[74] Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016.

[75] Cha Zhang, John C Platt, and Paul A Viola. Multiple instance boosting for object detection. In *Advances in neural information processing systems*, pages 1417–1424, 2006.

[76] James D Keeler, David E Rumelhart, and Wee Kheng Leow. Integrated segmentation and recognition of hand-printed numerals. In *Advances in neural information processing systems*, pages 557–563, 1991.

[77] Jan Ramon and Luc De Raedt. Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pages 53–60, 2000.

[78] Pedro O. Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[79] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.

[80] Ricardo Barandela, José Salvador Sánchez, Vicente Garcıa, and Edgar Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851, 2003.

[81] Kenneth O McGraw and Seok P Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.

[82] Chao Chen, Andy Liaw, and Leo Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110, 2004.

[83] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[84] Vincent C Cappendijk, Sylvia Heeneman, Alfons GH Kessels, Kitty BJM Cleutjens, Geert Willem H Schurink, Rob J Th J Welten, Werner H Mess, Robert-Jan van Suylen, Tim Leiner, Mat JAP Daemen, et al. Comparison of single-sequence t1w tfe mri with multisequence mri for the quantification of lipid-rich necrotic core in atherosclerotic plaque. *Journal of Magnetic Resonance Imaging*, 27(6):1347–1355, 2008.

[85] Norihide Takaya, Jianming Cai, MT Ferguson, S Marina, Vasily L Yarnykh, Baocheng Chu, Tobias Saam, Nayak L Polissar, Jane Sherwood, Ricardo C Cury, et al. Intra-and interreader reproducibility of magnetic resonance imaging for quantifying the lipid-rich necrotic core is improved with gadolinium contrast enhancement. *Journal of Magnetic Resonance Imaging*, 24(1):203–210, 2006.

[86] Bruce A Wasserman, William I Smith, Hugh H Trout, Richard O Cannon, Robert S Balaban, and Andrew E Arai. Carotid artery atherosclerosis: in vivo morphologic characterization with gadolinium-enhanced double-oblique mr imaging—initial results. *Radiology*, 223(2):566–573, 2002.

[87] Soumya Ray and Mark Craven. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd international conference on Machine learning*, pages 697–704. ACM, 2005.

[88] Ethem Alpaydın, Veronika Cheplygina, Marco Loog, and David MJ Tax. Single-vs. multiple-instance classification. *Pattern Recognition*, 48(9):2831–2838, 2015.

[89] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *arXiv preprint arXiv:1612.03365*, 2016.

[90] Gitte Vanwinckelen, Daan Fierens, Hendrik Blockeel, et al. Instance-level accuracy versus bag-level accuracy in multi-instance learning. *Data Mining and Knowledge Discovery*, 30(2):313–341, 2016.

[91] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[92] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[93] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[94] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

[95] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng Ann Heng. H-denseunet: Hybrid densely connected unet for liver and liver tumor segmentation from ct volumes. *arXiv preprint arXiv:1709.07330*, 2017.

[96] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.