VRIJE UNIVERSITEIT BRUSSEL

UNIVERSITEIT GENT

# Computer-aided diagnosis of ischemia and infarction for the treatment of acute ischemic stroke

Thesis submitted in order to be awarded the degree of Master of Science in Biomedical Engineering

Ine Dirks

Academic year 2017 - 2018

Prof. Dr. Ir. Jef Vandemeulebroucke
Dr. Koenraad Nieboer
Ir. Panagiotis Gonidakis

# Abstract

This thesis focuses on the treatment of patients suspected of suffering an acute ischemic stroke. This is a medical emergency in which there is a blockage in one of the brain vessels, reducing the blood flow to a certain region. Time is essential here because the longer it takes to restore the perfusion, the more brain tissue can suffer irreparable damage. Currently, the golden standard is to take a perfusion computed tomography and derive the perfusion parameters and maps from which a trained doctor can determine the severity and location of the problem. Based on this, certain treatment decisions can be made. A major disadvantage is that this is a complex procedure, only available at hospitals with a specialized stroke unit. When a patient has to be transported to the hospital, this may not be the nearest one. Since time is very valuable, this is not an optimal situation.

The aim of this dissertation is to analyze if the complicated derivation and interpretation of the perfusion maps can be replaced by a more straightforward method that can be used at any hospital. Machine learning is used to automatically detect the region of reduced perfusion and give an estimation of the severity. It is also investigated if the twenty scans taken in the present protocol are necessary to obtain all the required information. If this number can be lowered, there will be lower requirements imposed on the scanner, a decreased radiation dose for the patients and less time before the restoration of the blood flow.

Despite some limitations, a clear potential of using machine learning for this application is observed. Depending on the exact experiments, accuracies of 80 to 96 % are achieved. Moreover, our results show that a lower amount of scans can still provide a similar performance.

Further research is needed, including a larger and more diverse dataset and the use of more complex classifiers. However, this study confirms the potential to optimize the treatment of patients suffering an acute ischemic stroke and to offer them a chance of a better clinical outcome.

# Contents

# List of Abbreviations

| | |
|---|---|
| ACA | anterior cerebral artery |
| AIF | arterial input function |
| ANN | artificial neural network |
| ASL | arterial spin labeling |
| ASPECTS | Alberta Stroke Program Early Computed Tomography Score |
| CBF | cerebral blood flow |
| CBV | cerebral blood volume |
| CTA | computed tomography angiography |
| DCE | dynamic contrast enhanced |
| DSA | digital subtraction angiography |
| DSC | dynamic susceptibility contrast |
| DWI | diffusion-weighted imaging |
| FLAIR | fluid attenuation inversion recovery |
| GMM | Gaussian mixture models |
| IV | intravenous |
| MCA | middle cerebral artery |
| MIP | maximum intensity projection |
| MRF | Markov random field |
| MRI | magnetic resonance imaging |
| MTT | mean transit time |
| NCCT | non-contrast computed tomography |
| NIHSS | National Institutes of Health Stroke Scale |
| PCA | posterior cerebral artery |

| | |
|---|---|
| RBF | radial basis function |
| rLMC | regional leptomeningeal collateral |
| ROI | region of interest |
| SICH | symptomatic intracranial hemorrhage |
| SVM | support vector machine |
| TIA | transient ischemic attack |
| tPA | tissue plasminogen activator |
| TTP | time to peak |
| UZ | Universitair Ziekenhuis |

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Stroke

A stroke or cerebrovascular accident is a medical condition in which there is an inadequate blood supply to the brain, resulting in neurological symptoms that are often focal and acute. Worldwide, about 15 million people suffer form a stroke every year and it can happen to anyone at anytime [1]. As soon as the blood supply is disrupted, brain cells get damaged and there is little time before necrosis sets in. This can lead to permanent neurological damage and a drastic change in the patient's life.
In 80 % to 90 % of the cases a stroke is occlusive, meaning a blood vessel in the brain is blocked. This is called an ischemic stroke. Depending on the cause, it can be either thrombotic or embolic. A thrombotic stroke occurs because of a diseased or damaged artery that forms a blood clot in the brain. In case of an embolic stroke, the blood clot is formed outside the brain. Often it originates in the heart and travels through the vascular system until it's stopped in a cerebral vessel. Less occurring, 10 % to 20 % of the cases, is an hemorrhagic stroke, caused by the bleeding of a blood vessel. It can be either intracerebral, meaning the bleeding is located inside the brain, or subarachnoid, in which case there is a bleeding in the area immediately surrounding the brain.

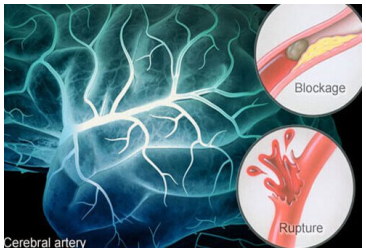



Figure 1.1: A stroke can be occlusive or hemorrhagic. Adapted from [2].

Possible first signs of a stroke are: speech impairment, confusion, reduced muscle strength at one side of the body, headache, vision impairment and drooping of one side of the face. In 1998, in the United Kingdom, the F.A.S.T. acronym was developed to help recognizing a stroke. Since then, it has been used and promoted by many stroke experts to inform people. The letters indicate possible symptoms of a stroke as illustrated in Figure 1.2.

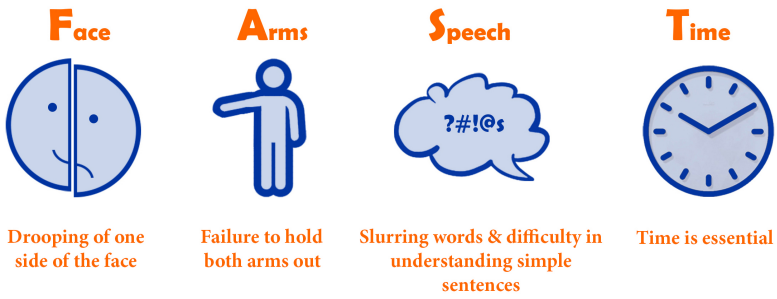



Figure 1.2: Typical signs of a stroke.

Another possible warning sign for a stroke is a transient ischemic attack (TIA). It causes symptoms similar to the ones of a stroke, but is resolved on its own within 24 hours or less. After this, the patient should be thoroughly tested and receive preventive therapy. With any kind of stroke there is ischemia, which diminishes the oxygen and glucose supply to the brain tissue and prevents the removal of waste products. Within seconds neurological signs can appear and within minutes there can be irreversible damage. Within the ischemic area, there are two major zones: the core and the penumbra, as illustrated in Figure 1.3. In the core, there is severe ischemia and the neurons and other cellular elements don't receive the oxygen and nutrients they need to survive. This results in infarction at which point the brain tissue cannot be saved anymore. The penumbra is a region around the core, with reduced blood flow, but that can still be salvaged if it is re-perfused in time. If the regained blood supply takes too long, this will gradually become part of the core. In brief, for the treatment of stroke, time is essential because 'time is brain' [3].

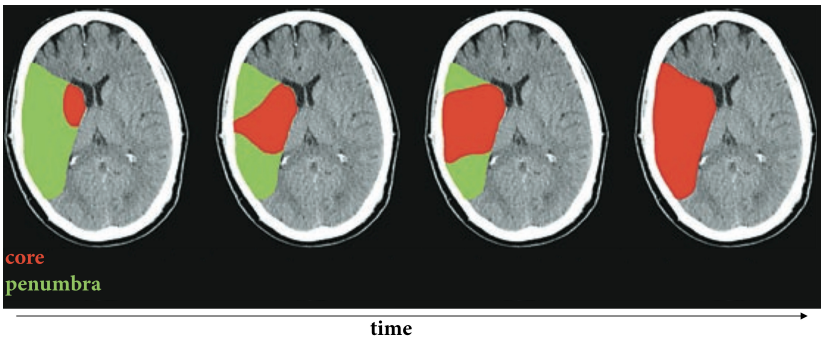



Figure 1.3: The evolution of penumbra and core over time. Adapted from [4].

In case of an arterial occlusion, some blood flow to regions distal to the blockage can be maintained thanks to alternative pathways, called cortical collaterals. They form a connection between larger cerebral arteries. Blood flow in function of the metabolic demand is possible thanks to the possibility for blood to flow in both directions. The better and more elaborate these collaterals, the better the blood flow will be in case of an occlusion. The strength and filling of these pathways can be used to predict the extent of the core and penumbra as well as the infarct growth. This way it can give an

indication for the choice of treatment. The status of the collaterals depends on numerous factors, like: genetics, age, prior medical conditions, etc. An illustration of the collateral circulation is shown in Figure 1.4.

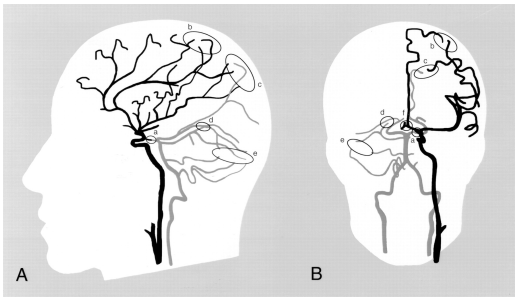



Figure 1.4: Arterial collateral circulation in lateral (A) and frontal (B) view. Adapted from [5].

## 1.2 Current treatment

Considering the penumbra, the tissue has to be re-perfused as fast as possible to avoid further infarction. First, it is essential to discover if the stroke is due to a thrombus or a hemorrhage because the treatment is very different. In case of a hemorrhage, intravenous (IV) thrombolysis has to be avoided since this will only cause more severe bleeding. To exclude the possibility of an hemorrhagic stroke, a non-contrast computed tomography (NCCT) scan is taken. If there is a bleeding, this shows up as a well-defined, hyperdense area, as is the case in Figure 1.5. This imaging technique is widely available in the hospital and offers a fast exclusion.

Figure 1.5: Non-contrast CT showing an intraparenchymal hemorrhage. Adapted from [6].

In case of an ischemic stroke, a choice has to made between IV-thrombolysis and thrombectomy. This is not straightforward and per patient an individual assessment of risks and benefits is needed to make a decision. With IV-thrombolysis, a bolus of tissue plasminogen activator (tPA) is administered intravenously. This will quickly dissolve the blood clot and restore normal blood flow. The time window for this method is 3 hours up to 4.5 hours [7]. However, there are certain contraindications for IV-thrombolysis, including

bleeding disorder, recent heart attack, recent head trauma, uncontrollable blood pressure, etc. The other option, thrombectomy, involves the surgical removal of the blood clot. In case of a large artery occlusion this offers a higher probability of a good clinical outcome [8]. In some cases a combination of both treatment methods is most adequate. For thrombectomy, further imaging is needed to localize the clot and analyze the status of the cerebral vessels. Also, to estimate the severity of the ischemia and infarction, a detailed imaging method is required.

## 1.3 Imaging options

Currently there are different imaging techniques available to assess the cerebral perfusion, including: conventional angiography, single phase computed tomography angiography (CTA), dynamic CTA, perfusion CT and magnetic resonance imaging (MRI). All of these have their advantages and disadvantages and will be discussed in the following paragraphs.

### 1.3.1 Conventional angiography

Conventional angiography requires the intravenous administration of a contrast agent. Through an X-ray this contrast agent, and thus the blood in which it is dissolved, is visualized. The quality can be further improved through image processing. Often, a baseline scan is taken before the administration of the contrast agent. This can be subtracted from the X-ray with contrast to enhance the visibility of the blood vessels, in which case it is termed digital subtraction angiography (DSA). Taking multiple scans during the passage of the contrast agent can give the temporal information needed to asses the blood circulation. This technique has a good image quality and allows to visualize the blood vessel originating from one specific artery. However, it is a 2D imaging method that is time consuming. It is more suitable to guide a thrombectomy than to visualize the entire cerebral vasculature.

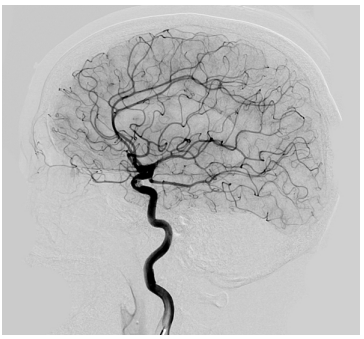


Figure 1.6: Conventional angiography. Adapted from [9].

### 1.3.2 Single phase CT angiography

CT angiography follows the same principle as the conventional angiography, but instead of a planar X-ray, a CT scan is taken, acquiring a 3D representation of the cerebral blood vessels. In single phase CTA, one scan is taken at a certain moment in time, so

it does not include any time information. Therefore, the timing for taking the scan is crucial, often the collateral status is mislabeled and it can't represent blood circulation. To eliminate the influence from the background, again subtraction imaging is possible.

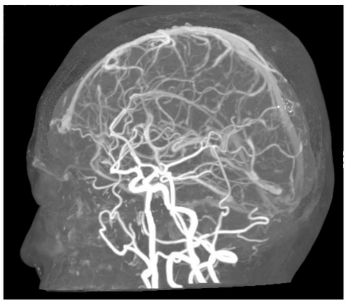


Figure 1.7: Maximum intensity projection from a single phase CTA. Adapted from [10].

### 1.3.3 Dynamic CT angiography

Dynamic CTA consists of multiple single phase CTA's taken at different points in time. It contains the necessary temporal information to asses the blood circulation. Angiography is important to review the status of the vessels and check for any elements that could hinder the thrombectomy.

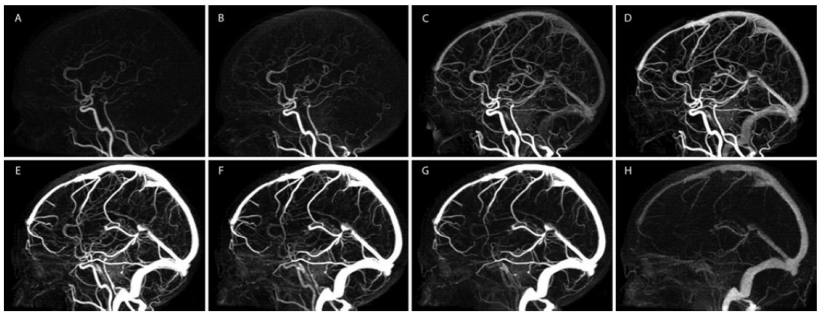



Figure 1.8: Maximum intensity projections from dynamic CTA. Adapted from [11].

### 1.3.4 Perfusion CT

Perfusion CT or dynamic contrast-enhanced (DCE) CT is currently the golden standard for assessing blood distribution in the brain. A contrast agent is injected and a series of CT scans is taken at different time points to image the cerebral blood circulation. From these images, the hemodynamic characteristics can be derived. The contrast anywhere in the tissue ($C_{tissue}$) has to be compared to the contrast in the artery ($C_{artery}$) to minimize any uncertainties about the bolus characteristics. The total amount of contrast at the artery is called the arterial input function (AIF). The cerebral blood volume (CBV) at any position can be estimated by

$$CBV = \frac{\int_{t=0}^{\infty} C_{tissue}(t)\,dt}{\int_{t=0}^{\infty} C_{artery}(t)\,dt}. \tag{1.1}$$

The blood is flowing from the artery to the tissue with a certain speed. There is a temporal relationship that can be defined as

$$C_{tissue}(t) = C_{artery}(t) * h(t). \tag{1.2}$$

where $*$ represents the convolution between both concentrations and $h$ the transit times. These times are dependent on the amount of contrast still present after a certain time $t$. This situation is described by the residue function,

$$R(t) = 1 - \int_{\tau=0}^{t} h(\tau)\,d\tau. \tag{1.3}$$

From this, the relationship between the contrast in the artery and the contrast in the tissue follows.

$$\begin{aligned} C_{tissue}(t) &= CBF\,(C_{artery} * R)(t) \\ &= CBF \int_{\tau=0}^{t} C_{artery}(\tau)\,R(t-\tau)\,d\tau, \end{aligned} \tag{1.4}$$

with $CBF$ the cerebral blood flow, i.e. the amount of blood passing by per unit time. The amount of time it takes the blood to pass through the vascular bed is referred to as the mean transit time ($MTT$). It can be calculated by dividing the blood volume by the blood flow or taking the area under the residue function.

$$MTT = \frac{CBV}{CBF} = \int_{0}^{\infty} R(t)\,dt. \tag{1.5}$$

Another important perfusion parameter is the time to peak ($TTP$), the time for the contrast in the tissue to reach its maximum concentration. The time for the residue function to reach its peak is referred to as $T_{max}$.
All of these perfusion parameters are illustrated in Figure 1.9.

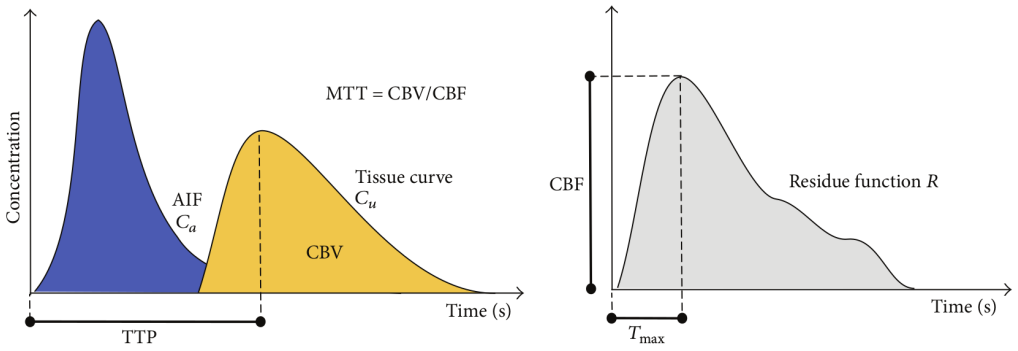
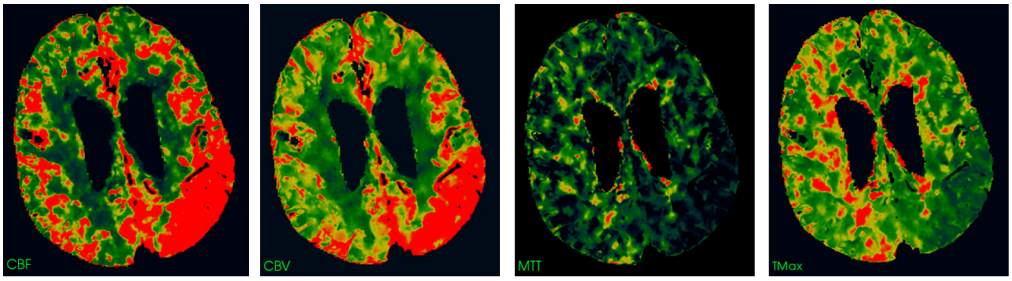



Figure 1.9: Perfusion parameters. Adapted from [12].

The contrast in the tissue and the AIF can be derived from the intensities of the CT scans. However, the residue function and the blood flow are unknown, meaning there will be more than one possible solution to the equations. There are different techniques for estimating the perfusion parameters, divided into two main categories: compartmental analysis [13, 14] and deconvolution analysis [15]. The exact method will depend on the software that is used. Once these perfusion parameters are derived per voxel, the perfusion maps are constructed, showing the parameters for every voxel in the brain. An example is shown in Figure 1.10.




(a) (b) (c) (d)

Figure 1.10: Perfusion maps: cerebral blood flow (a), cerebral blood volume (b), mean transit time (c) and TMax (d).

For a trained doctor, these maps give a clear understanding of the blood perfusion in the brain tissue and help to assess the location and severity of the ischemia. Using these perfusion parameters, segmentations of core and penumbra can be made based on certain thresholds. However, there is no consensus in literature about the optimal perfusion parameters and thresholds that should be employed to get as close to reality as possible [16–20]. The resulting ischemic tissue maps will depend on which software is used. An example of such classification is presented in Figure 1.11. The perfusion parameters and derived maps are important for the diagnosis of a stroke, estimating the location and severity of the ischemia and for making decisions regarding the treatment. The downside of perfusion CT is that about 15 to 20 minutes [21] are needed to take the series of scans and interpret the results. It is a complex procedure that requires trained personnel and a recent scanner. Currently, it is only available in specialized hospitals.

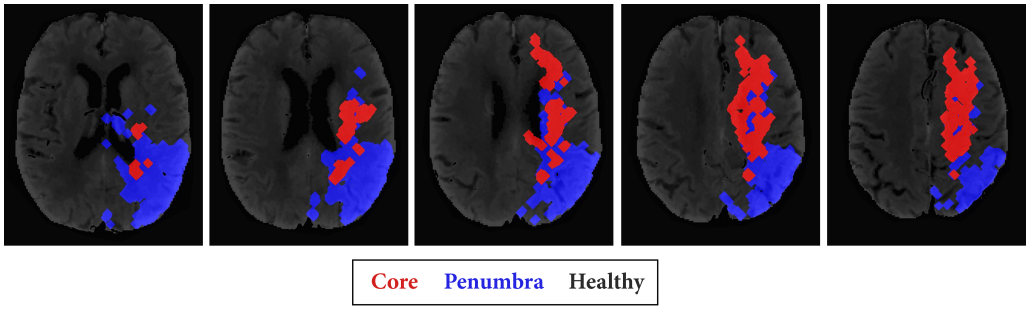



Figure 1.11: A few slices showing an example of ischemic tissue segmentation

### 1.3.5 Magnetic resonance imaging

MRI provides the most contrast in soft tissue, allowing a more accurate depiction of infarction in the brain compared to CT. Yet, the acquisition and interpretation can take more than 20 minutes [21], which is too long for the treatment of stroke patients. However, it can be used during the patient follow-up after treatment.
There is a wide variety of acquisition methods to obtain the wanted contrast. Dynamic contrast enhanced MR imaging and dynamic susceptibility contrast (DSC) imaging make use of a contrast agent to increase the visibility of blood on the MRI. DCE employs T1-weighted images to evaluate the perfusion towards the extravascular, extracellular space. The passing contrast agent will temporarily increase the signal. DCS, on the other hand, requires T2*-weighted images to look at the blood vessels where the contrast agent will attenuate the signal.
Diffusion-weighted imaging (DWI) uses a T2*-signal to evaluate how water molecules can diffuse. If they can move around more easily, there will be a loss of signal, resulting in a more black region on the image. But when the water cannot diffuse easily, in an area of reduced perfusion for example, the signal will be higher, showing a brighter spot on the image.
Another option in MRI perfusion imaging is arterial spin labeling (ASL). Here, a reference image and labeled image are acquired. For the latter, a radiofrequency pulse is applied that saturates water protons to magnetically label the water molecules in the arterial blood. The only difference between both images is the magnetization of this incoming blood. When they are subtracted, the static signals cancel out. No external contrast agent is needed since magnetically labeled blood acts as a tracer.
Although MRI offers some interesting properties, CT is preferred over MRI. A CT only takes a few seconds, while MRI can take several minutes. An additional advantage of CT is that it is widely available in the hospital, especially in an emergency department.

## 1.4 Current challenges

In the treatment of stroke, every 30 minutes delay can increase the risk of poor clinical outcome by 14 % [22]. This makes CT the preferred imaging technique since it only takes a few seconds and is widely available in the hospital. The current golden standard, perfusion CT, is a complex method requiring fast scanning capabilities, perfusion analysis software and experienced personnel to perform the acquisition and interpretation. Currently about twenty CT scans are taken in a perfusion protocol. This requires a recent scanner, with fast acquisition capabilities and sufficient coverage and exposes the patient to a relatively high radiation dose. When a patient suspected of having a stroke is transported to a hospital, this might not be the nearest one. Since time is essential in case of stroke, this is not an optimal situation.
There is a need for a method that can provide the same information needed to take a therapeutic decision, but that can be performed with any CT scanner, requires less complicated processing and more straightforward interpretation. Being able to transport the patient to the nearest hospital and not the nearest one with a stroke unit can reduce the amount of time to start the treatment. This can improve the clinical outcome and have a big impact on the patient's life after the stroke.
A lot of research has been conducted towards optimizing the stroke treatment protocol. The most relevant studies are discussed in the next chapter.

# Chapter 2

# State of the art

## 2.1 Image acquisition

The acquisition method for imaging the brain is an important aspect in the treatment of stroke. It should give accurate information, that can be used to make treatment decisions, as fast as possible. Several studies [23–27] showed that the status of collateral blood vessels is a good indicator for the clinical outcome after acute ischemic stroke.
Martinon et al. [28] performed a literature review studying different acquisition methods to evaluate collateral circulation. Fifteen studies were described using angiography to assess the cortical collaterals. However, this method is not optimal in case of stroke because of the additional time it requires. Also, the conclusions about the state of the collateral circulation strongly depend on the timing and speed of the scans.
Another fifteen cases were discussed using CTA. This modality is easier and faster than conventional angiography. However, it is still not ideal to image the cerebrovascular circulation because it provides no temporal information. To overcome the previous issues, two other methods are being researched: dynamic CTA and multiphase CTA.

### 2.1.1 Collateral circulation

#### Predict the response to IV-thrombolysis

Calleja et al. [25] provided a method to determine a score for the leptomeningeal collateral arteries from perfusion CT source imaging to predict the response to IV-thrombolysis. In patients with ischemic stroke, a good collateral score resulted in a better response to IV-thrombolysis. They showed better clinical outcome, both short- and long-term, as well as a smaller final infarct volume.

#### Collateral status

The research of Menon et al. [29] assessed the collateral status from dynamic CTA. For this, they also used the hemodynamic information that is only available by this type of acquisition. From the images, they identified the retrograde filling of pial arteries distal to the occlusion.
The results showed a high variability in the rate of backfilling in pial arteries which could be a marker of the collateral status. However, further research is needed to make con-

clusions regarding this relationship.

#### Single phase CTA vs. dynamic CTA

The study of Frolich et al. [30] compared single phase CTA with dynamic CTA and with volumetric perfusion maps. They used the regional leptomeningeal collateral (rLMC) score [31] which varies between 0, meaning there is no visible collateral flow, and 20, when the collateral flow is equal to or greater than the flow in the unaffected hemisphere. The study showed that dynamic CTA has an improved sensitivity for characterizing the collaterals compared to conventional CTA. Also, time-resolved maximum intensity projection (MIP) images are best to look at the collateral status and to predict clinical outcome. For the latter, it is best to look at the total extent of collateral flow which means that the amount of visible collaterals should be maximized in the used imaging modality.

#### Collateral status in relation with infarct volume and follow-up

The research of van Den Wijngaard et al. [32] compared dynamic with single phase CTA, focusing on the collateral status in relation with the infarct volume and follow-up. This showed that dynamic CTA offers the possibility for a more detailed assessment of the collaterals while the use of single phase CTA results in an underestimation of the collateral score. This way, when using dynamic CTA, there is a stronger relation with the infarct core during follow-up and the obtained collateral scores can be used for a more accurate prediction of clinical outcome. They also found that, for the prediction of the evolution of the infarct volume, the optimal time point for acquisition of the collaterals of the affected hemisphere is after the peak arterial phase of the unaffected hemisphere.

### 2.1.2 Multiphase CT angiography

Another option is to reduce the amount of scans in dynamic CTA to obtain multiphase CT angiography.

#### Acquisition

A description of the acquisition is provided by Menon et al. [22]. After the NCCT, the cerebral blood vessels are visualized at three different phases: peak arterial phase, equilibrium or peak venous phase and late venous phase. This is illustrated in Figure 2.1. In the first phase, a conventional CTA is obtained from the aortic arch to the skull vertex. The next two images span from the skull base to the vertex.
By reducing the amount of scans, both time and scan dose are lowered. Also, this technique requires no additional contrast agent, it provides whole-brain coverage and needs no post-processing. This technique proved to be a quick and reliable alternative to dynamic CTA. However, it also showed some limitations. For example, proximal stenosis, collaterals in the base of the skull or poor cardiac function may lead to a delay in contrast flow and mislabeling of the collateral status. This might be resolved by a fourth imaging phase. Moreover, this protocol cannot be used in patients with posterior circulation stroke because of poorly understood hemodynamics [22].

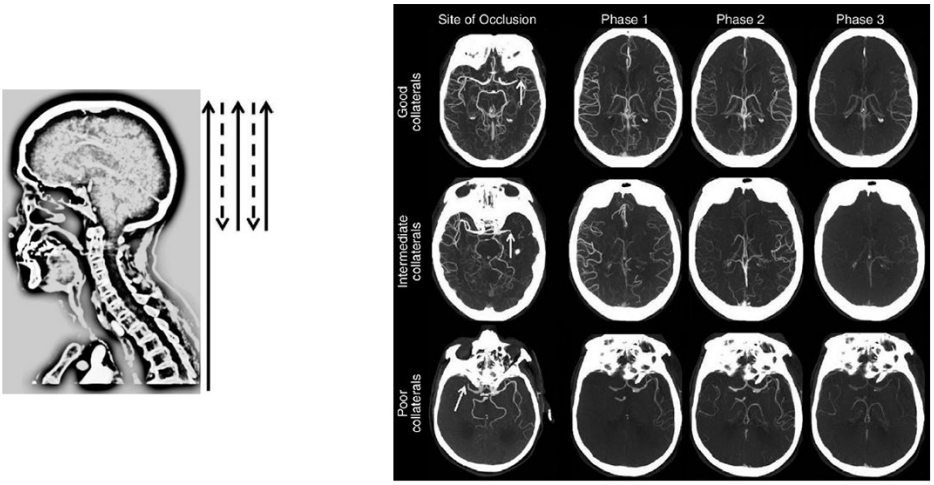



Figure 2.1: Multiphase CTA acquisition (left), the dashed arrows represent the movement of the scanner in between scans, and obtained images (right). Adapted from [22].

### Effectiveness

Recently, more studies have been conducted regarding this method by Yu et al. [33], D'Esterre et al. [34] and Volny et al. [35]. All of them proved that multiphase CTA is an adequate and reliable method for examining the collateral circulation in stroke patients. The work of Yu et al. [33] confirmed that multiphase CTA increases the diagnostic accuracy compared to single-phase CTA. Furthermore, they included non-radiology trained physicians and trainees in the study to prove that multiphase CTA is an imaging method that is easy to use. They could indeed rapidly and accurately indicate anterior circulation occlusions which is an important measure in the revascularization procedure.

### Multiphase CTA vs. perfusion CT

D'Esterre et al. (2017) [34] compared multiphase CTA with perfusion CT in predicting the clinical outcome of tissue after acute ischemic stroke. From multiphase CTA following parameters are considered: delay in maximal pial vessel enhancement compared with the contralateral hemisphere, washout of contrast within pial vessels and extent of maximal pial vessel enhancement compared with the contralateral hemisphere. Their ability to predict the clinical outcome of tissue is checked and compared to the quality of perfusion CT parameters. This showed that both imaging modalities have comparable accuracy in predicting the evolution of the infarction. Moreover, the parameters of the multiphase CTA have similar values to the CBF, CBV and Tmax of the perfusion CT. For multiphase CTA, the washout corresponds best to the evolution of the infarction. It shows the amount of local perfusion pressure. A longer washout time means poor collateral supply and thus severe ischemia. The corresponding perfusion CT parameter for this is the Tmax. The delay in multiphase CTA corresponds to CBF and the assessment of pial vessel filling can be the new parameter to indicate the ischemia status and the potential core tissue in function of the re-perfusion time. An important limitation of this study is that only occlusions of the M1 segment of the middle cerebral artery (MCA) are

considered. Further research is needed to take into account more distal blockages.

#### Multiphase CTA vs. single phase CTA

Also in 2017, Volny et al. [35] made a comparison between single phase and multiphase CTA, focusing on occlusions in the M2 segment of the MCA. Their results showed that the sensitivity for distal or secondary clot detection is significantly better in multiphase CTA. Both well and less experienced radiologists were included. The agreement in the less experienced group was a lot better when using multiphase CTA which proved to be a quick and user-friendly tool. Another big advantage compared to perfusion CT is that it can be implemented in most conventional scanners. Limitations of the study include a small sample size of 20 cases and the acquisition using both methods was performed in different patients.

#### More research

More studies [24, 36, 37] can be found regarding the multiphase CTA, all resulting in similar conclusions. Although all of the previous studies had their limitations, it is clear that multiphase CTA is a reliable tool for the assessment of collateral circulation in patients with acute ischemic stroke. Yet, it is too early to conclude that multiphase CTA can simply replace perfusion CT with its quantitative perfusion maps.

## 2.2 Prediction of clinical outcome

### 2.2.1 Based on collaterals

The research mentioned in the previous section aimed at finding the most adequate way to image and characterize collaterals in order to use this information for predicting the clinical outcome of the patient.
Recent research by Tong et al. [38] takes this a step further. They developed a software that can assess the collaterals through machine learning and score them to predict the clinical outcome for the patient. The software provides an assessment method that is objective, quantitative and semi-automatic. Perfusion CT scans are used to obtain MIP images and automatically segment brain vessels. Next, these are categorized based on their origin using machine learning. This is possible by assuming that the time-intensity curve of a certain vessel is most similar to the curve of the artery it emerges from. For this, there is first a training phase where a training set for the machine learning algorithm is derived. This training set consists of the time-intensity curves of following regions of interest (ROI):

- segment A1 for the anterior cerebral artery (ACA)
- segment M1 for the middle cerebral artery
- segment P1 for the posterior cerebral artery (PCA)

These ROIs are shown in Figure 2.2.

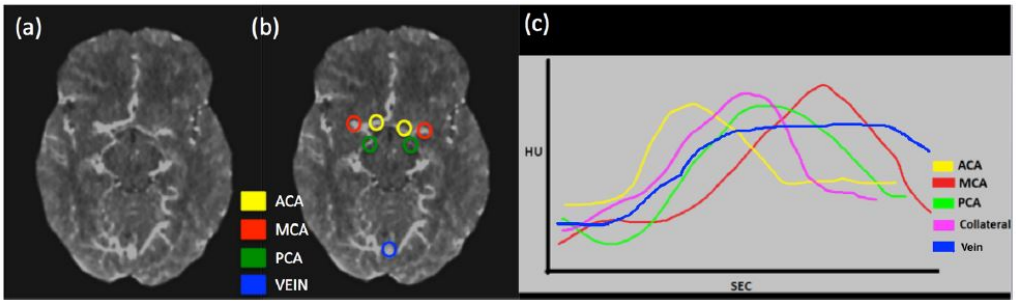



Figure 2.2: Maximum intensity projection image (a) with regions of interest (b) and corresponding time-intensity curves (c). Adapted from [38].

Using principle component analysis, a multi-dimensional space is derived from the training set, characterized by a set of eigenvectors. Transferring the training waveforms into this space, each curve is represented by a unique linear combination of these eigenvectors with a weighting coefficient, the eigenvalue.
The second step is the recognition phase. Here, the remaining brain vessels are classified depending on their origin. For each vascular pixel, the time-intensity curve is derived and projected onto the multidimensional space giving a certain set of eigenvalues. By minimizing the distance between these values and the ones from the training set, the origin can be found. By counting the number of pixels belonging to each of the three vascular territories, a total volume of collaterals is obtained and used for determining the collateral score. This results in a high collateral score for vessels receiving blood from mainly the ACA and PCA. A low collateral score corresponds to collaterals mainly supplied by the MCA, which is fully or partially occluded. Through a multivariate logistic regression model they tested the prognostic value of age, baseline National Institutes of Health Stroke Scale (NIHSS), Alberta Stroke Program Early CT Score (ASPECTS), recanalisation status and collateral score. This assessment method proved to have a prognostic and predictive value. However, there are some limitations to the study. The number of patients was small (135), there was limited coverage (4 cm in craniocaudal direction) and two injections of contrast agents were used.

### 2.2.2 Based on the entire CT

Another study aimed at predicting the outcome of thrombolysis through the use of machine learning on CT brain images [39]. More specifically, the model would predict if the procedure would benefit the patient or cause a symptomatic intracranial hemorrhage (SICH). They trained a first SVM using an entire CT scan, postprocessed according to their own pipeline, combined with the NIHSS. A second, more user-dependent SVM was constructed by inputting the NIHSS together with parameters regarding the extent of the ischemia determined by a radiologist. The performance of the models was assessed by 10-fold cross-validation. They concluded that the SVM can separate the patients that will benefit from thrombolysis from the ones that will develop a SICH because of it. Moreover, the automated method resulted in a more accurate performance than the typical, radiologic methods. However, the study did have important limitations. Only a small number of the considered patients actually suffered a SICH. Consequently, there is no guarantee the model is general enough to perform well on just any patient. Another

restriction is their use of non-enhanced CT images which are not the optimal modality to detect hemorrhage related features. A final consideration mentioned is that their image processing technique was inefficient and took about 30 minutes per scan. Despite these challenges, the study shows the potential of using machine learning in stroke treatment.

## 2.3 Automatic prediction

Automatic prediction of the core and penumbra in case of a stroke might offer an alternative to the complex derivation of perfusion parameters and maps. Currently, to the best of our knowledge, prediction of core and penumbra directly from CTA or perfusion CT, has not been reported. Related approaches have however been proposed for various kinds of tissues on data from different imaging modalities.

### 2.3.1 High-grade glioma on MRI

Zikic et al. [40] developed a method for automatic segmentation of six-channel MR images in case of high-grade gliomas. These tumors consist of edema and the gross tumor, which is made up of active cells and a necrotic core. For the classification of the tissue, decision forests are used, made up of decision trees. They're combined with a generative model of tissue appearance by using probability estimates based on Gaussian mixture models (GMM) as extra input for the decision forest. The decision trees operate in a training step and a testing step.
Three different context-aware feature types are used:

1. intensity difference between a point in one channel and its offset point in another channel

2. difference between intensity means of a cuboid around a point in one channel and around its offset point in another channel

3. intensity range along a 3D line between a point and its offset point in the same channel

The decision tree is made up of nodes, containing a set of training examples and a class predictor. This class predictor is the probability for the number of points with a certain class. At the start of the decision tree, all spatial points in all training data sets are taken into account. At every node the training examples are split up according to their feature representation. This process is stopped at a certain depth.
Next, during the testing phase, the point that has to be categorized is pushed through each tree according to the acquired split functions. When the point arrives at a leaf node, this probability is used as the tree probability for that point. The overall probability is determined as the average of all tree probabilities. The class the point belongs to is then the most probable class.
The method from Zicik et al. [40] proved to be a very accurate segmentation procedure, as can be appreciated from Figure 2.3. The method is computationally efficient and fairly robust to parameter settings. From the obtained classifications it is possible to do volume measurements that can be indicative for treatment planning.

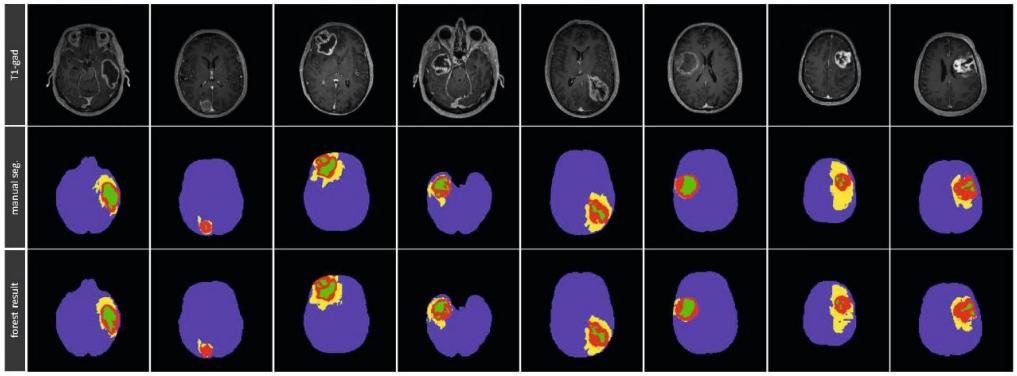



Figure 2.3: Examples of results on eight patients, adapted from [40].
T1-weighted image, post gadolinium (top row), manual segmentation ($2^{nd}$ row)
and the random forest prediction (bottom row).

### 2.3.2 Ischemic stroke on MRI

A method for automated segmentation of ischemic tissue after stroke is developed by Mitra et al. [41]. Multi-channel MR images were acquired three months after the stroke: T1-weighted, T2-weighted, fluid attenuation inversion recovery (FLAIR) and DWI. The work flow is sketched in Figure 2.4.
After pre-processing, a first step is the hierarchical segmentation of probable lesion from FLAIR images. From the intensities of these images, a first assumption of five classes is made: background, gray matter, white matter, cerebrospinal fluid and lesion. Next, they are further classified into eight classes. Then, three classes with the highest means are combined, forming one probable binary lesion class, corresponding to stroke and white matter lesions. The classification is done through Bayesian inference and Markov random field (MRF) segmentation. Next, an expectation maximization algorithm is applied to obtain a lesion and non-lesion class. The likelihood of this lesion class is then used as an input for the random forest training. The images obtained from the different processing steps are summarized in Figure 2.5. The Bayesian and MRF segmentation gives a good approximation of the lesion area. Going from the initial five-class segmentation (Figure 2.5 (b)) to the eight-class (Figure 2.5 (c)) improves the classification of the lesion and its heterogeneities. Figure 2.5 (d) shows the result of combining three classes into the probable lesion class to obtain a binary image. Figure 2.5 (e) shows the lesion class likelihoods after the eight-class expectation maximization algorithm while Figure 2.5 (f) indicates the probabilities obtained from the random forest calculations. It is clear that the probabilities from the likelihood are noisy while the random forest offers smoother areas and improved segmentations. Comparing the final result to the ground truth segmentations made by an expert (Figure 2.5 (g)), validates the quality of the segmentation method.
This study demonstrates that using likelihood of lesion areas as an input for a random decision forest improves the obtained segmentations. A limitation of the study is that it cannot quantify the tissue loss specifically due to ischemic stroke and there are no separate classes for ischemic, white matter and secondary lesion areas.

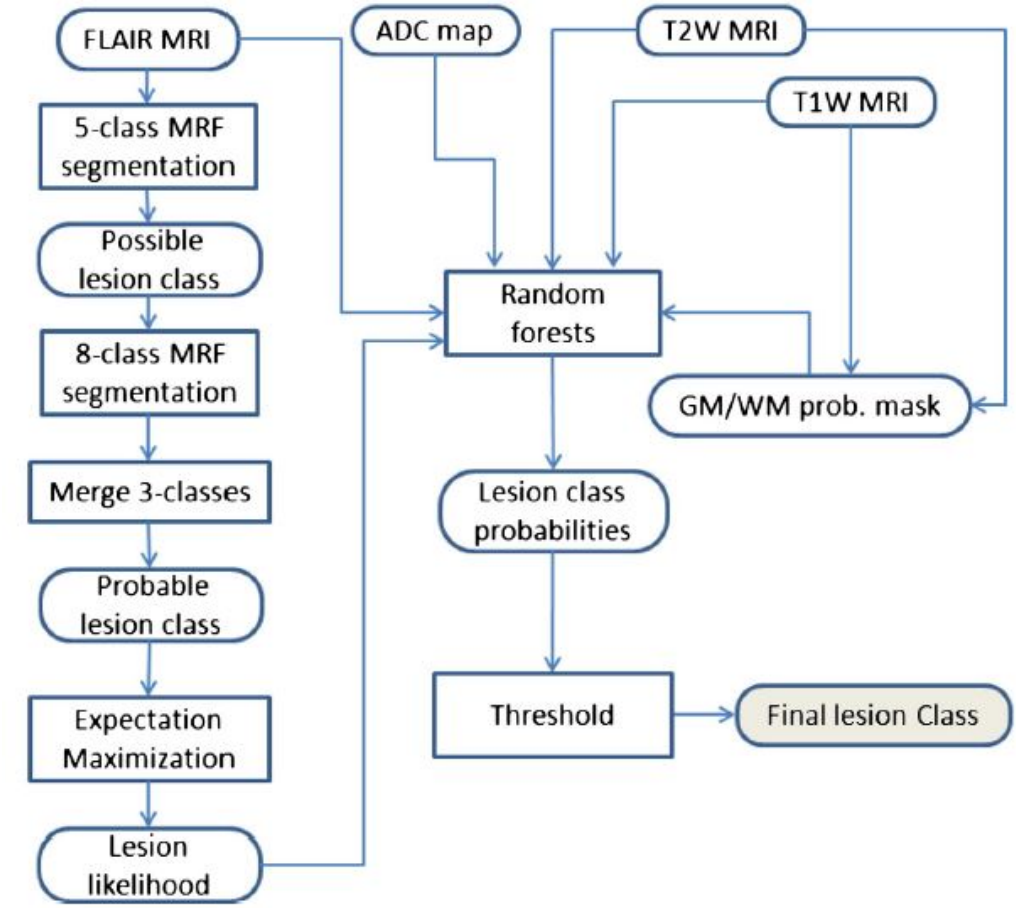



Figure 2.4: Work flow of Mitra et al. [41].

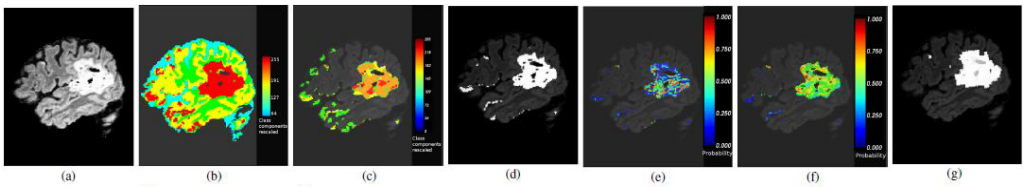



Figure 2.5: original FLAIR image (a), 5-class Bayesian and MRF segmentation (b), 8-class Bayesian and MRF segmentation leading to three classes (c), binary of the three combined classes (d), likelihood of the probable lesion, maximizing the probabilities in the lesion area when divided into lesion and non-lesion (e), random forest probabilities (f), ground truth segmentation made by an expert (g). Adapted from [41].

In 2017, the paper of Lee et al. [42] was published, presenting an overview of recent applications of machine learning as an aid in the diagnosis and prognosis of stroke. For ischemic strokes, three studies are mentioned on automatic lesion segmentation [43–

45], but these are using MRI and DWI instead of CT.
In the work of Maier et al. [43], a comparison is made between nine different classifiers:

- K-nearest-neighbors
- Gaussian naive bayes
- Generalized linear models
- Gradient boosting
- Adaboost
- Random decision forest
- Extra tree forest
- Tuned extra tree forest
- Convolutional neural networks

MR images are used with as features: their intensity, the weighted local mean, the 2D center distance and the local histogram. The training and testing set were preprocessed by downsampling them, performing intracranial segmentation, bias field correction and intensity standardization. The obtained segmentations were postprocessed by removing islands smaller than 1.5 ml. These were assumed to be the result of some noise that was still present or of potential errors induced by the automatic preprocessing pipeline. The best performances were realized by the random forest and the convolutional neural network, proving this is a complex problem, requiring an advanced machine learning method. Besides the different classifiers, the manual segmentations from two experienced observers were compared to obtain an inter-observer score. Although some of the classifiers performed well, none of them achieved results in the range of the human agreement. An example of their results is shown in Figure 2.6.

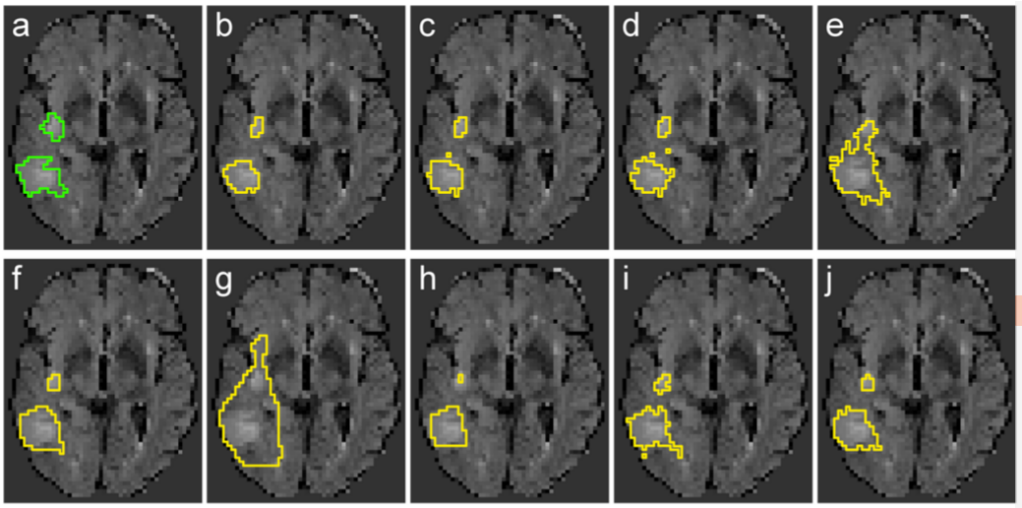
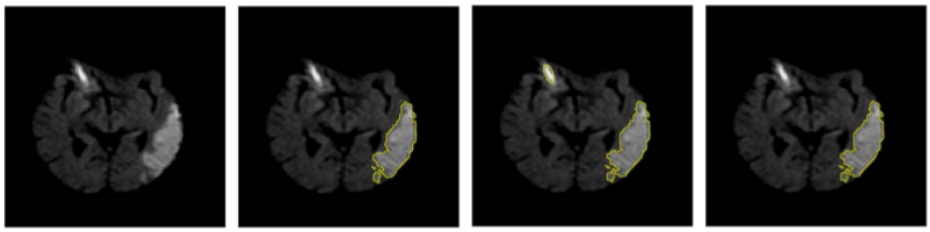



Figure 2.6: Example of results by Maier et al. [43]. Ground truth (a), 100 nearest neighbors (b), 10 nearest neighbors (c), 5 nearest neighbors (d), AdaBoost (e), extra trees (f), Gaussian naive bayes (g), generalized linear models (h), gradient boosting (i) and random decision forest (j).

Pustina et al. [44] proposed a new method to automatically segment stroke lesions from T1-weighted MR images, called 'Lesion Identification with Neighborhood Data Analysis' (LINDA). From these images, they used geometric features, atlas-based deviation,

control-based deviation and subject-specific anomalies. The classifiers were multiple random forests, each trained with different image resolutions. The outcome of each random forest was used as additional input for the next one, creating successive improvements of the prediction. The model included 60 cases with a stroke in the left hemisphere. Testing was done by k-fold cross-validations, by leave-one-out evaluations and by applying the model on data from other institutions, all yielding good results for a large number of patients.
Chen et al. [45] looked into convolutional neural networks to segment stroke lesions on DWI. They propose a system composed of two modules. The first one consists of N adapted DeconvNets and is called the 'EED net' which has to segment the lesions. The second part, the MUlti-Scale Convolutional Label Evaluation Net (MUSCLE Net), targets the small detected lesions and tries to remove the false positives. They obtained good results, however there is a margin for improvement regarding different sizes of lesions, as well as false negatives. Figure 2.7 shows an example of a segmentation.



(a) (b) (c) (d)

Figure 2.7: Example of a segmentation: original-diffusion weighted image (a), manual reference (b), EDD result (c) and EDD + MUSCLE result (d). Adapted from [45].

### 2.3.3 Hemorrhagic stroke on CT

In the paper of Lee et al., only one study is mentioned, by Scherer et al., where CT images are used for automatic lesion segmentation [46] and this is for hemorrhagic stroke, not ischemic. They used a random forest classifier for the automatic segmentation. Per voxel, local gray value image features were used, characterizing the neighborhood of the voxel. Volume labeling for intracerebral hemorrhage, subarachnoid space and parenchyma was added by manual segmentations. After training and testing, some post-processing steps were taken on the prediction maps, including smoothing and morphological operations. The model was trained and tested through 5-fold cross-validation and finally tested on an independent validation sample. The study showed a strong agreement with the manual segmentations.

## 2.4 Goals of this thesis

As mentioned in the previous chapter, there is a need for a stroke treatment protocol that can be used in every hospital. Also, there might be potential to optimize the number of scans needed, leading to a wider availability of the treatment procedure and a lower radiation dose for the patient.
A lot of research has been focusing on analyzing the collateral circulation as a predictor for the clinical outcome of patients after acute ischemic stroke and after certain therapies. A recently developed imaging technique, multiphase CT angiography proved to be a good method to visualize these collaterals and draw reliable conclusions. It needs less time than perfusion CT, less post-processing, gives a lower scan dose to the patient, is more user-friendly and can be incorporated in the majority of currently available scanners. However, at the moment, the multiphase CT is still used in combination with perfusion CT. Each of the studies had their limitations and more research is needed for further developments in this direction.
Some other studies looked into automatic lesion prediction by the use of machine learning on CT or MRI. However, the ones targeting ischemic stroke all used MRI images to automatically determine the core lesion. In stroke treatment, CT is preferred over MRI because it is faster and more available at the hospital. At the time this dissertation was written, no studies were found using machine learning algorithms on perfusion CT images for the automatic segmentation of core and penumbra.
The goal of this thesis is to develop a method to asses the state of the cerebral perfusion in stroke patients that can be used in any hospital and still provides the necessary information for the doctor to make the therapeutic decisions. It is investigated if the complicated derivation and interpretation of perfusion parameters and maps can be replaced by a machine learning algorithm that predicts the size and location of the penumbra and core, using perfusion CT images as input. Also, an analysis is made regarding the amount of scans that are necessary to provide the doctors with all the required information. If there is a reduction possible in the number of scans without losing the necessary information, this can lead to a wider availability thanks to the lower requirements for the scanner hardware as well as a diminished scan dose for the patient.

# Chapter 3

# Materials and methods

## 3.1 Machine learning algorithms

Machine learning is a subdivision of artificial intelligence that uses statistical methods to allow computers to learn from data. There are three main types: supervised, unsupervised or reinforcement learning. Here, only the first one will be used. In supervised learning the data are provided together with the desired outputs and the model learns by example. When then new, unseen data are inputted, the model makes a prediction based on the rules it learned from the initial dataset. There is a wide variety of machine learning methods, each with their own characteristics. In the following paragraphs the techniques used in this thesis are explained. For all the machine learning, the Python library $Scikit-learn$ is used.

### 3.1.1 Support vector machine

Support vector machine (SVM) can be used as a binary classifier that searches for the hyperplane or set of hyperplanes that maximizes the margin. This is the distance between the plane and the closest datapoints, called the support vectors. In Figure 3.1 there are three support vectors for the blue class and two for the orange class.

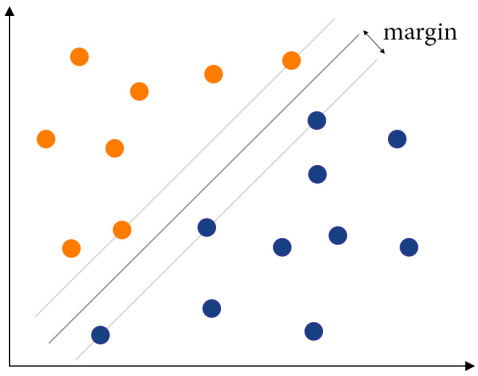



Figure 3.1: Principle of a support vector machine.

By maximizing the margin, the model will be as general as possible and perform best on new, unseen data. This technique works well on data that is linearly separable. However, often the dataset is more complex and cannot be classified with a linear SVM. In that case a method called the 'kernel trick' can be implemented. The original feature space is then mapped to a higher dimension where it is separable again, as illustrated in Figure 3.2.

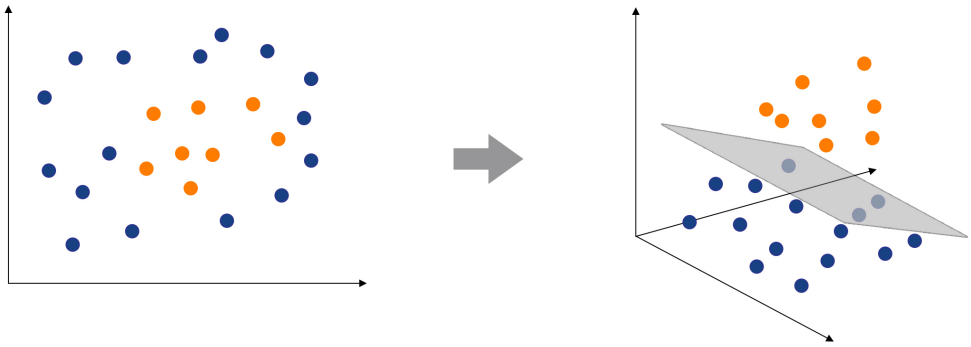
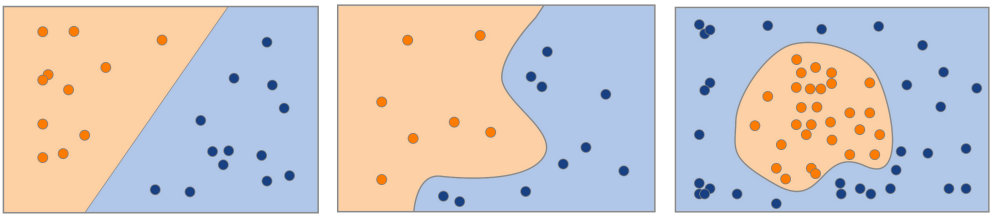


Figure 3.2: Principle of the kernel trick.

Different kernels can be used to go to a higher dimension. Some common ones are: linear kernel, polynomial kernel and radial basis function (RBF). They differ in how they define the decision boundaries as shown in Figure 3.3.



(a) (b) (c)

Figure 3.3: Difference between the decision boundaries of a linear kernel (a), a polynomial kernel (b) and a radial basis function (c).

Using SVM, there are some parameters that can be set to tune the model and find a balance between fitting to the training data and generalizing the model. It is important to avoid overfitting. This means the model performs well on the training data, but not on unseen data because it corresponds too closely to the training set. There are two important parameters that can be set for this tuning: $C$ and $\gamma$.
$C$ defines the penalty for misclassification of a sample. When $C$ is small, there will be more errors in classification, but also a smoother decision boundary. If $C$ is high, the model tries to classify all training samples correctly, leading to an irregular boundary.
The parameter $\gamma$ is imposed for the polynomial and RBF kernels. It determines the influence of a training sample on the decision boundary. If $\gamma$ is low, this impact will be far, leading to an edge without a lot of curvatures. When $\gamma$ is high, the influence will not reach that far and the boundary is more curved. This can create islands of classification

areas around one or a few samples.
Although SVM is a binary classifier, it can also be used for problems with more than two classes. By combining multiple SVM's, the dataset can be separated into several subsets.

### 3.1.2 Decision tree

A decision tree is a machine learning method that can classify a dataset in multiple classes. It typically starts from one node branching into different possible outcomes. Each of these form nodes that can branch again into several possibilities. This continues until either all training samples are classified or one of the limiting parameters, set before training, is reached. Figure 3.4 shows a simplified scheme of a decision tree.

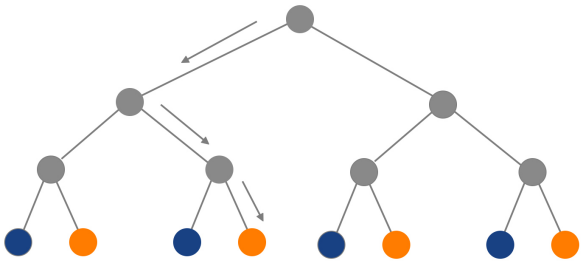


Figure 3.4: Principle of a decision tree.

Every node creates a new, linear decision boundary. For example if node 1 checks if feature $x_1$ is greater than 10, this creates a boundary at $x_1 = 10$. Next, node 2 might separate the samples based on feature $x_2$ by testing if these are greater or smaller than 15. This example is shown in Figure 3.5.

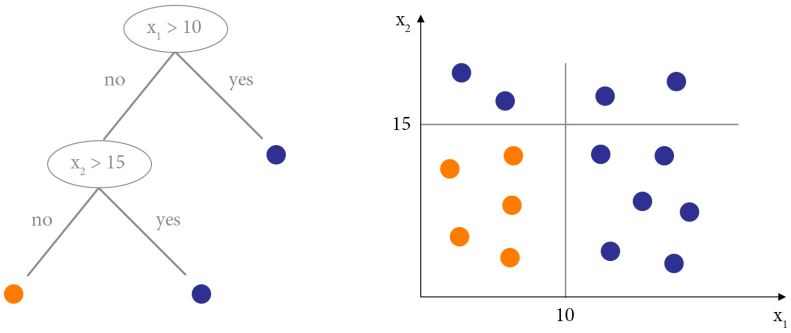



Figure 3.5: Example of a classification by a decision tree.

In reality, the datasets and decision trees are much more complex, but the principle stays the same.
Again, different parameters can be tuned to find the optimal balance between fitting on the training data and acquiring a generalized model. Important ones are the maximum depth and the minimum number of leaf samples. The maximum depth can limit to how far down the tree will branch off. If the depth is too high, there is a high chance of

overfitting. If it is too small, there might not be enough decision boundaries to model a complex dataset. The minimum number of leaf samples determines how many elements have to be in one leaf node. This will also influence the depth of the tree.

### 3.1.3 Random forest

A random forest combines of a number of decision trees using the bagging method. This means random dataset samples are taken from the original set, with replacement. A number of classifiers are then trained on this subset. If these make random, uncorrelated errors, the performance will be improved because the influence of these errors is reduced. In addition, when splitting a node, the best split among a random sample of features is taken instead of all the features. The forest is constructed by growing different trees for each group of samples and features. Compared to one decision tree, a forest introduces a randomness that will slightly increase the bias, but decrease the variance thanks to averaging. The latter will usually dominate, leading to an overall better performance. Figure 3.6 presents a simplified scheme of the random forest principle.

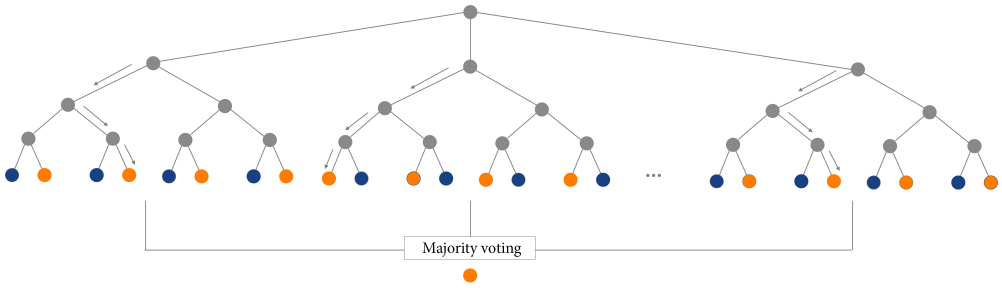



Figure 3.6: Principle of a random forest.

### 3.1.4 Deep learning

Deep learning is a subdivision of machine learning that can extract relevant features by itself and use these for training. It can also construct more complex decision boundaries, without being restraint by linear ones. This is achieved by constructing multiple layers where each one uses the input from the previous one. This can lead to a very complex artificial neural network (ANN), that is capable of handling complicated problems. Figure 3.7 sketches this principle.

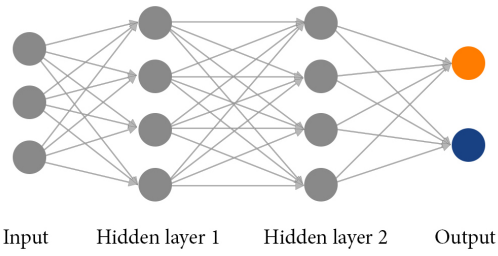



Figure 3.7: Principle of an artificial neural network.

Since this is a first attempt in using machine learning on these data, it was chosen to first test machine learning techniques other than deep learning, with the aim to characterize the data and enhance our understanding of this classification problem. However, it is important to acknowledge its potential and when used correctly, it can offer greater results then the classifiers mentioned above.

### 3.1.5 Voxel-wise classification

For the prediction of core and penumbra, every voxel in the image containing brain tissue is a separate sample that can be used for training. Each of these voxels has a certain set of features. For example, from a perfusion CT these features can be the intensities of that voxel at each point in time. Besides these features, the actual tissue class is also given as input. Once the model is trained on this set, a new image can be tested. For every voxel, with its set of features, the trained model will assign a tissue class. This principle is sketched in Figure 3.8.

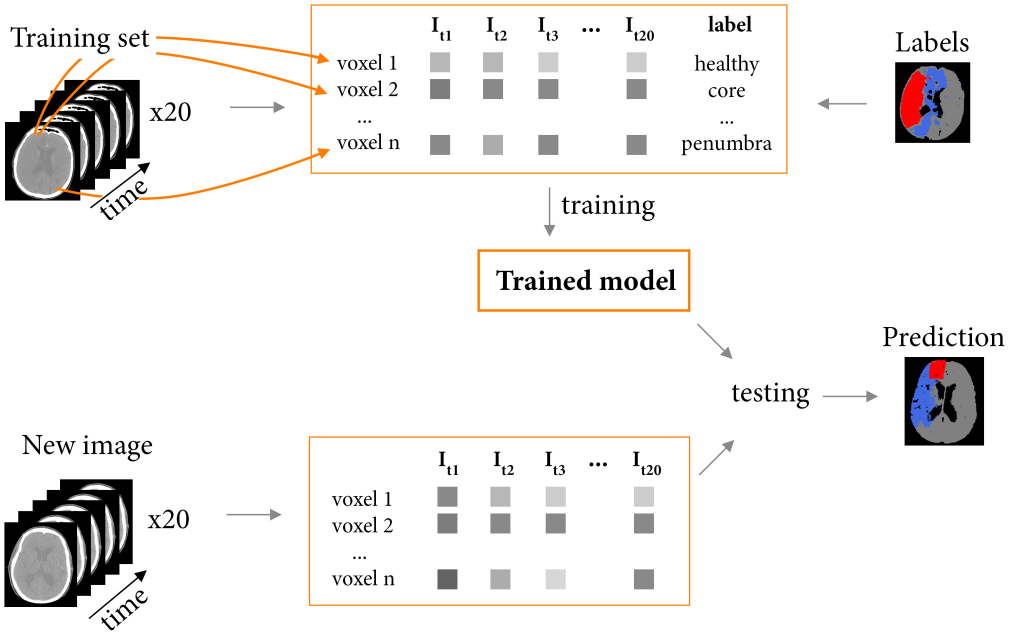



Figure 3.8: Principle of the voxel-wise classification.

## 3.2 Data properties

The data consist of images from 44 patients treated at the Universitair Ziekenhuis (UZ) Brussel, including: a NCCT, a multiphase CTA and a perfusion CT. An overview of these scans and their characteristics is given in Table 3.1.

| | **NCCT** | **Multiphase CTA** | | **Perfusion CT** |
|---|---|---|---|---|
| | | **Phase 1** | **Phases 2 and 3** | |
| Size x [# voxels] | 512 | 512 | 512 | 512 |
| Size y [# voxels] | 512 | 512 | 512 | 512 |
| Size z [# voxels] | 512 | 1121 – 1333 | 512 – 484 | 16 / 24 |
| # scans in time | 1 | | 3 | 20 |
| Spacing x [mm] | 0.40 – 0.48 | 0.39 – 0.52 | 0.39 – 0.52 | 0.35 – 0.40 |
| Spacing y [mm] | 0.40 – 0.48 | 0.39 – 0.52 | 0.39 – 0.52 | 0.35 – 0.40 |
| Spacing z [mm] | 0.31 | 0.31 | 0.31 | 5 |
| Slice thickness [mm] | 0.625 | 0.625 | 0.625 | 5 |
| Peak voltage [kV] | 120 | 120 | 120 | 80 |
| Tube current [mA] | 468 – 492 | 149 – 264 | 149 – 264 | 240 |

Table 3.1: Scan properties.

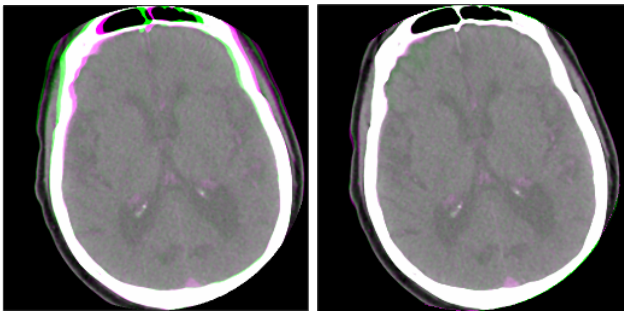
The dataset consists of twenty healthy patients with normal cerebral perfusion. The remaining 24 patients have locally reduced blood perfusion, showing a penumbra and sixteen of them also have a core region. These include patients with stroke mimics and patients that actually suffered from a stroke.

## 3.3 Data preprocessing

To yield an optimal performance of the machine learning algorithms, the data has to be preprocessed. This includes: registration, filtering, baseline subtraction, normalization, whitening, balancing and randomizing. All of these steps are explained in the next paragraphs.

### 3.3.1 Image registration

The perfusion CT and multiphase CTA consist of a number of scans taken at different points in time. It is possible the patient moves during this procedure, causing a misalignment between the different images. For the voxelwize classification it is important that the same voxels at different timepoints can be selected. Therefore, the images at all timepoints are registered to the one of the first scan. In Figure 3.9, a comparison is made between the situation before and after registration. This is an overlay of the first and last scans where the green and purple indicate the difference between both images. Besides the perfusion CT, also the NCCT is registered to this first image to be able to use both scans together later on.

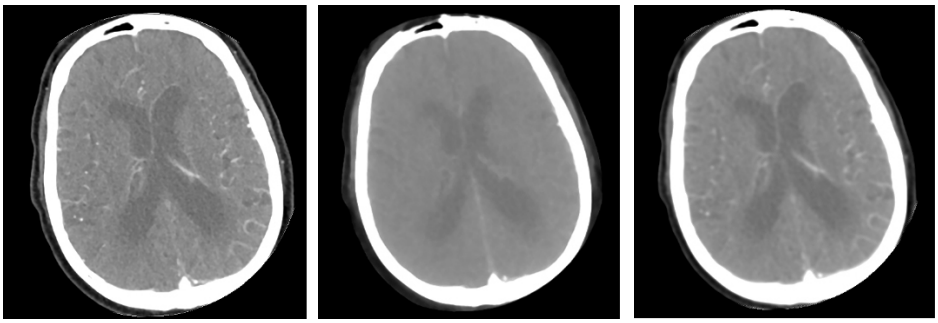


(a) (b)

Figure 3.9: Overlay of the first and last scan in a perfusion CT: before registration (a) and after registration (b).

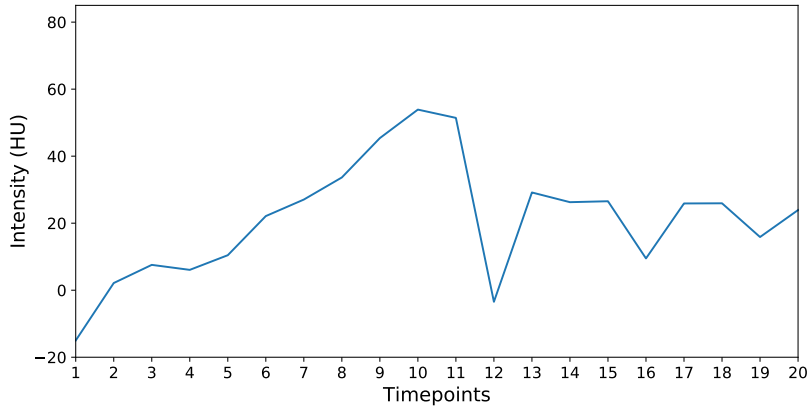
### 3.3.2 Image filtering

There is a lot of noise present in the images, as can be observed from the perfusion curve in Figure 3.11 (a). Therefore, an important step is filtering. For perfusion images, it is crucial that the injected contrast is not filtered out. A method preserving the essential features is needed. A common median filter will only preserve the edges, but smooth out the outliers, which in this case is also the contrast agent. A better option is the anisotropic diffusion filter [47]. The principle is that a set of gradually more blurred images is created by the convolution between the image and a 2D Gaussian filter. The filter parameters depend on the local content of the image, making it a space-variant transformation. It is an iterative process where the next image of the set is obtained by transforming the previous one.
A comparison between the effect of a median filter with kernel size 3 and a anisotropic diffusion filter with conductance 3 and 5 iterations is made in Figure 3.10. It is clear that the latter is the best option for these data. The effect is clarified by the perfusion curve in Figure 3.11 (b).

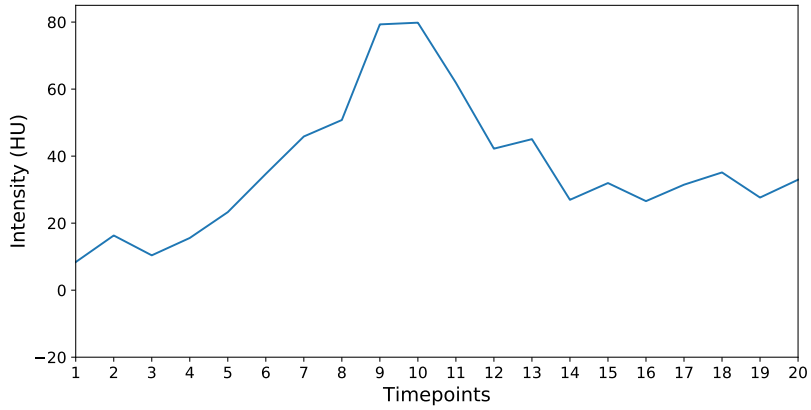


(a) (b) (c)

Figure 3.10: Filtering the data: no filter (a), median filter (b) and anisotropic diffusion filter (c).




(a)

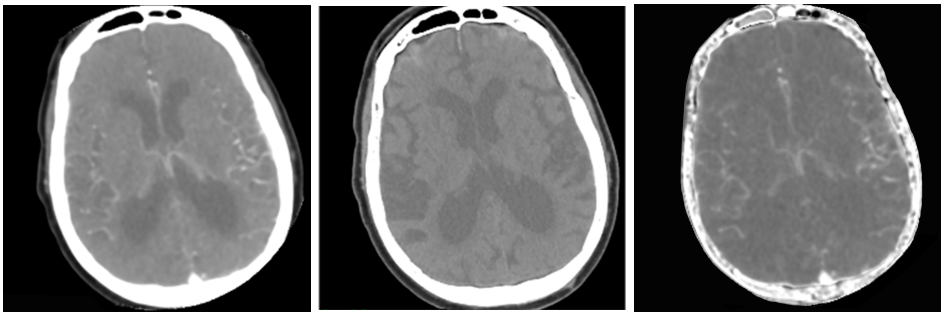



(b)

Figure 3.11: Perfusion curves of one voxel showing the effect of filtering the image: initial curve (a) and curve after anisotropic diffusion filtering (b).

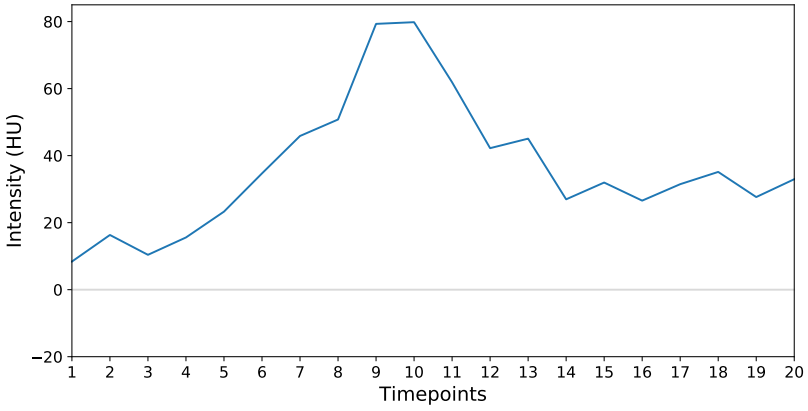
### 3.3.3 Baseline subtraction

The determination of the perfusion parameters is based on concentrations of contrast in the tissue. We hypothesize that contrast concentrations will constitute better features for the machine learning classification than absolute intensities in HU. Therefore, the baseline is subtracted from the perfusion CT. In the stroke imaging procedure, the first scan is always a NCCT to exclude any bleeding. This will be used here as baseline to subtract from the perfusion CT. From here on, anytime the intensities are mentioned, this means the baseline subtracted intensities, unless specified otherwise. This image preprocessing step is illustrated in Figure 3.12 and the effect on the perfusion curve can be observed from Figure 3.13.

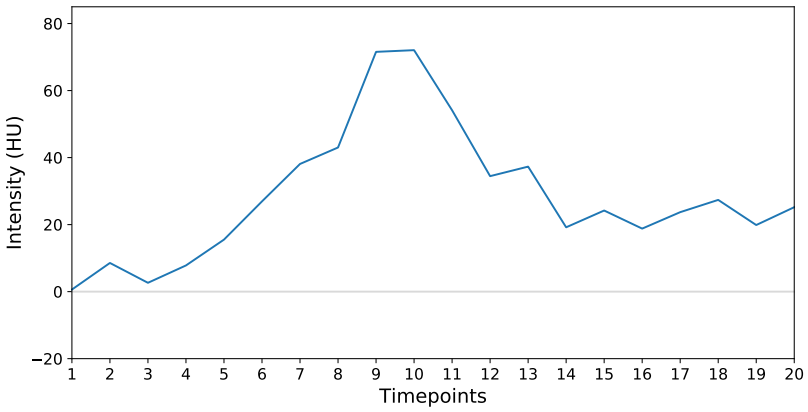


(a) (b) (c)

Figure 3.12: Illustration of baseline subtraction: initial image with contrast (a), baseline image (b) and baseline subtracted image (c).




(a)




(b)

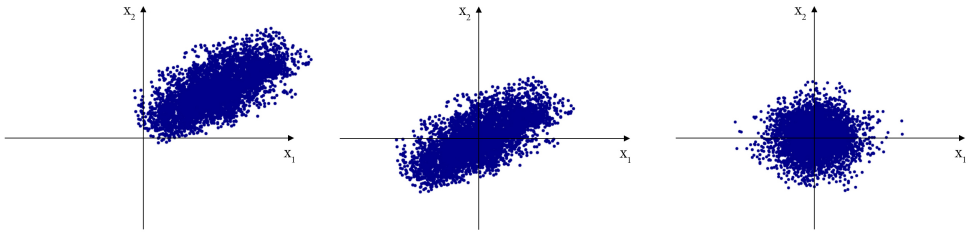
Figure 3.13: Perfusion curves of one voxel before (a) and after (b) baseline subtraction.

### 3.3.4 Normalizing and whitening

Most machine learning methods perform best if the features are normalized, so if they have zero mean and unit variance. This can be obtained by z-scoring, according to the equation

$$z_n = \frac{x_n - \bar{x}}{\sigma}, \tag{3.1}$$

where $x_n$ is the original feature value, $\bar{x}$ is the mean over all samples for that feature and $\sigma$ is the standard deviation.
In addition, whitening can also improve the performance of the classifier. This is a linear transformation applied on data with zero mean and a known covariance matrix in order for this to turn into the identity matrix. This way the data will be uncorrelated.
The effect of these preprocessing steps is illustrated in Figure 3.14.




(a) (b) (c)

Figure 3.14: Illustration of preprocessing steps: initial data (a), data after z-scoring (b) and data after z-scoring and whitening (c).

### 3.3.5 Balancing

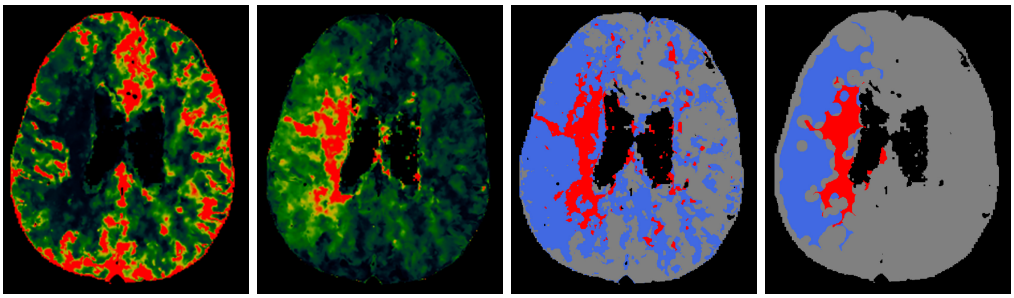
To avoid bias towards one class, the dataset can be balanced, obtaining an equal amount of samples for each class. This can be done by downsampling the classes with a higher number of samples by taking a random subset with the same size as the smallest class. Another option is to upsample the smaller classes. Here, synthetic samples are created based on the ones from the training set.
Both balancing approaches will be tested to check which one yields better results.

### 3.3.6 Randomizing

A final step in the preprocessing of the data is randomization of the samples. Some machine learning algorithms are sensitive to the observation order of data. It is good practice to randomize this order and not have a high number of successive examples of the same class.
Whenever a randomization is performed, a random state is used, introducing a seed value. This initializes a pseudo-random number generator, making sure the same random sample is taken every time, this to ensure results of different experiments can be compared. Randomization is performed as a final preprocessing step, but is also part of downsampling a dataset, using a random forest, etc. In every step where some kind of randomization happens, a seed value is included.

## 3.4 Ground truth labels

The ground truth labels for the three classes are derived from the perfusion maps. As mentioned before, there is no consensus in literature about the optimal perfusion parameters and thresholds to determine core and penumbra [16–20]. This thesis is about analyzing whether or not machine learning can offer an alternative to the complicated process of perfusion analysis by predicting core and penumbra segmentations based on the perfusion CT. Finding the optimal parameters and thresholds for the labels to come as close as possible to reality, is a study on its own and is out of the scope of this thesis. Therefore, the ground truth labels are developed by using parameters that give a reasonable segmentation for all patients. These include, for penumbra, a $T_{max}$ value of more than 6 seconds and, for the core, a CBF of less than 3.5 ml/100g/min. There is a lot of noise present in the perfusion maps, leading to some dispersed labeling of voxels and non-uniform segmentations. In reality, a region of reduced perfusion is likely to be interconnected because of the vasculature in the brain. Therefore, it is safe to assume the labels should form a segmentation that is connected in the three dimensions. To make the labels more homogeneous, morphological operations are used. This includes a dilation, followed by erosion to close the binary masks. Next, only the largest connected region over three dimensions is retained to obtain a final label, not including scattered patches of voxels of a certain class. The results of these steps are illustrated by an example in Figure 3.15.



(a) (b) (c) (d)

Figure 3.15: Example of ground truth labels: cerebral blood flow map (a), TMax map (b), ground truth labels (c) and postprocessed ground truth labels (d)

# Chapter 4

# Feasibility study

The images available for this thesis have never been used for machine learning applications before. This is why this chapter comprises an initial prestudy, performed to check the feasibility of using machine learning algorithms on these data. Three common approaches, SVM, decision tree and random forest, are tested on interpatient basis, using a simplified dataset generated from manually selected voxels. The aim is to verify the initial performance of different classifiers which could then be employed in full scale, realistic studies.
The classifier performing best is then used on a more realistic dataset, containing all the voxels in the brain. These experiments are performed on an intrapatient basis, i.e training and testing data was obtained from the same patient scan, as the limited patient data available at the time did not allow to develop a general model.

## 4.1 Manual selection of data

These first experiments are performed to get a good understanding of the data and the performances of the different machine learning methods. For this, training and testing are performed on different patients, but the voxels are selected manually to make sure they are representative for the classes. Also, only two classes are considered: core and healthy. For now, the penumbra is not included because this would make the classification more complicated. An example of the voxels that were selected is shown in Figure 4.1, together with their time-intensity curves. The red one represents the AIF, the blue one is healthy tissue while the orange is core. From the graphs, it is clear that these are good examples for each of the classes. There is indeed more contrast flowing in the healthy tissue than in the core.
To obtain enough samples to feed into the algorithms, the neighboring voxels of the ones that were selected, are included. An area of 20 on 20 voxels is considered, centered around the chosen voxel. Since the spacing in the z-direction is quite large, 5 mm, only samples within the same slice are considered. This results in 400 voxels per class, so 800 voxels for one patient. For training, six patients are included while one is kept aside for testing. So, in total the training dataset consists of 800 x 6 = 4800 samples, of which 2400 are healthy and 2400 are core. The features consist of the baseline-subtracted intensities at 20 timepoints and each of the three classifiers is tested with a wide range of parameters. The performance is reviewed by two different metrics. First, the average accuracy from a 10-fold cross validation is determined by the *cross_val_score* function of Scikit-learn.

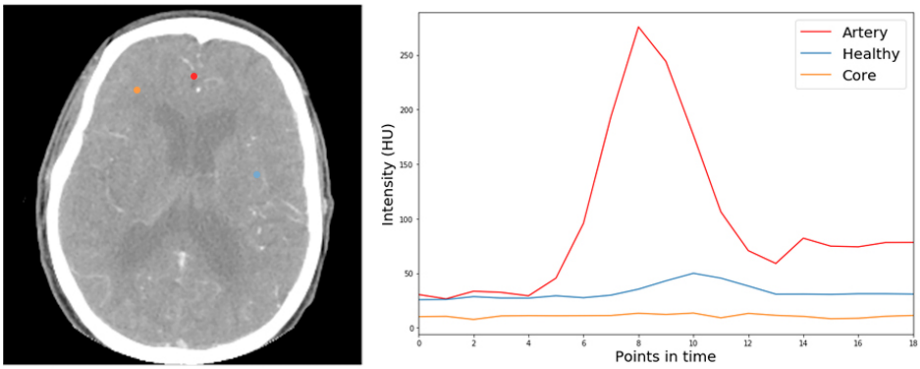



Figure 4.1: Voxels selected for the first, simplified experiment and the corresponding time-intensity curves.

This splits the training set into 10 subgroups. The model is trained on 9 of them and evaluated on the remaining set. The accuracy is defined as

$$accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i), \tag{4.1}$$

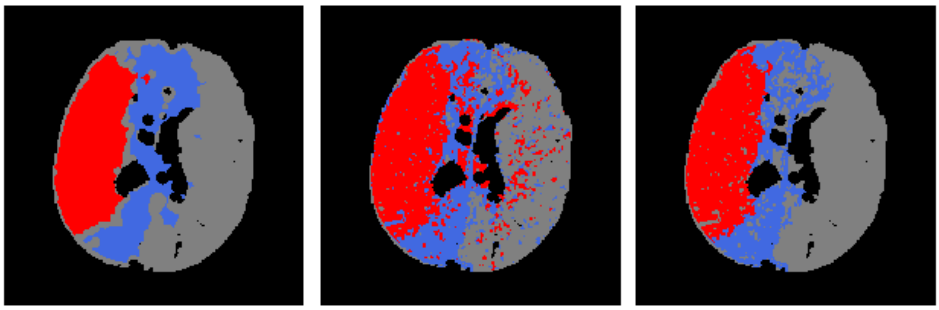
where $y_i$ is the true value of the $i$-th sample and $\hat{y}$ is the corresponding predicted value, $n$ is the number of samples and $1(\hat{y}_i = y_i)$ is the indicator function, consisting of 0 for all elements that are different in $y$ and $\hat{y}$ and 1 for all identical elements. This is repeated ten times, each repetition using a different set for testing. From the ten resulting accuracies, the mean is taken to obtain one, final score. This method represents how well the classifier separates the training voxels, regardless from which patient they are derived.
The second metric determines the accuracy on the test set, the patient's dataset that was not included in the training. This is assessed using the *accuracy_score* function of Scikit-learn where the accuracy is defined the same way as in equation 4.1. This evaluation is repeated seven times, each time keeping another patient aside for testing. This way, a leave-one-out cross-validation is performed. It is the most important metric since in reality the scan from a new patient will be inputted to predict the ischemic zones. However, for now, it is interesting to compare both metrics to better understand the performance of the classifiers. If, at a certain point, the difference between both accuracies increases, this can indicate that the model is overfitting to the training data and some parameters should be adjusted. An overview of all results is given in Appendix A.1 while the best results per classifier are summarized in Table 4.1.
The highest achieved, average accuracy through a leave-on-out experiment on seven patients using SVM is 82.95 %. Using a decision tree this is 82.03 % and using a random forest this is 84.87 %. These results prove that there is potential to use machine learning to predict the tissue classification from a perfusion CT scan. However, this experiment is performed for a simplified situation with only a small number of samples, that were assured to be representative for both tissue classes. A next experiment is performed, including the entire brain and the three classes: healthy, core and penumbra. Since the random forest performed best, this classifier will be used for all further predictions.

| Parameters | | Test patient | Accuracy cv* on training set | Accuracy on test set |
|---|---|---|---|---|
| **SVM** | | | | |
| C = 100 | $\gamma$ = 0.01 | 1 | 0.9992 | 0.8213 |
| | | 2 | 0.9985 | 0.9900 |
| | | 3 | 0.9990 | 0.8813 |
| | | 4 | 0.9988 | 0.6688 |
| | | 5 | 0.9992 | 0.8225 |
| | | 6 | 0.9992 | 0.8650 |
| | | 7 | 0.9992 | 0.7575 |
| **average** | | | **0.9990** | **0.8295** |
| **Decision tree** | | | | |
| Max depth* = none | MLS* = 10 | 1 | 0.9502 | 0.7000 |
| | | 2 | 0.9408 | 0.9688 |
| | | 3 | 0.9460 | 0.9250 |
| | | 4 | 0.9592 | 0.7325 |
| | | 5 | 0.9571 | 0.8313 |
| | | 6 | 0.9477 | 0.8850 |
| | | 7 | 0.9608 | 0.7000 |
| **average** | | | **0.9517** | **0.8204** |
| **Random forest** | | | | |
| N = 200 | | 1 | 0.9744 | 0.8463 |
| | | 2 | 0.9733 | 0.9888 |
| | | 3 | 0.9767 | 0.9338 |
| | | 4 | 0.9858 | 0.7288 |
| | | 5 | 0.9779 | 0.7888 |
| | | 6 | 0.9763 | 0.9025 |
| | | 7 | 0.9792 | 0.7525 |
| **average** | | | **0.9776** | **0.8488** |

Table 4.1: Results from manually selected points in the brain for the three classifiers with values for the parameters that gave the best performance (*cv = cross-validation, max depth = maximum depth, MLS = minimum number of leaf samples, N = number of estimators).

## 4.2 Incorporation of the whole brain

In reality, voxels from the entire brain need to be considered, including the ones where the classification might not be that straightforward. Also, besides healthy and core, a third class is introduced: penumbra. Not to complicate the procedure too much, these tests are performed on intrapatient basis and not between different patients. The classification problem is quite complex and at the time this experiment was conducted, insufficient data was available to obtain a generalized model. Because the purpose of this chapter is only to check the feasibility of this thesis, intrapatient analysis is sufficient for now.
From the seven patients, there are four that present all three classes of brain tissue. From these, only one has a core large enough to keep enough training samples compared to testing samples after balancing the set by downsampling. For this patient, one slice is left out as test set while training is performed on the remaining slices. A random forest is used with 200 estimators, no restriction on the depth and minimum 10 leaf samples as this yielded the best results in the previous experiment.
The accuracy on the test slice is 75.54 % and the resulting classification is provided in Figure 4.2 (b). Compared to the ground truth, Figure 4.2 (a), the model clearly recognizes the areas with reduced perfusion and can make a distinction between core and penumbra. The sizes of both regions are also comparable to the ground truth. There are, however, some small patches of voxels with the wrong classification, distributed over the slice. In reality, all regions in the brain are connected by the cerebral vessels. If there is ischemia due to a blockage, the area with reduced perfusion is not likely to comprise several smaller, scattered fragments. Therefore, from the prediction, only the largest connected area per class is kept. This is similar to what was done in the study of Maier et al. [43]. The result of this step is shown in Figure 4.2 (c). Now the accuracy is increased to 82.69 %. The performances on the other slices of this patient are similar and tabulated in Appendix A.2.



(a) (b) (c)

Figure 4.2: Prediction of one slice by a random forest, trained on the remaining slices of that patient: ground truth (a), prediction (b) and postprocessed prediction (c).

## 4.3 Conclusion

Since this is the first time machine learning is applied to the images used in this thesis, a feasibility study was performed.
A first experiment was completed on a simplified dataset, derived from seven patients, to get an understanding of the data derived from a perfusion CT, the performance of different classifiers and the complexity of the problem. Good results were obtained with accuracies from 82 % to 99 %. Although the SVM and decision tree also achieved good results, the random forest performed best and will be used for the further predictions.
Next, an entire brain was considered, containing less representative samples as well as the penumbra class. For this, the experiment was performed on intrapatient basis and not between different patients. The model could never be general enough to yield good results with only seven patients since blood perfusion is a complex process with parameters that differ between people. One slice was taken out as test set, while the model was trained using the remaining slices of that patient. Also here, promising results were obtained that could be further improved using postprocessing.
Although, only a simplified version of the problem was considered, the potential of using machine learning to replace the complicated derivation of perfusion parameters and maps, was indicated. In the following chapters, the dataset can be expanded and more complex experiments can be conducted.

# Chapter 5

# Predicting core and penumbra

In this chapter the dataset of seven patients is expanded to 44 cases. From these, there are sixteen containing the three classes: healthy, core and penumbra. Eight only have a penumbra and the remaining twenty patients are completely healthy with respect to the cerebral perfusion.
Furthermore, additional features are tested as well as the reduction of the number of scans in time.

## 5.1 Feature extraction

### 5.1.1 Intensities as features

In reality, the images from a new, unseen patient will be analyzed to predict the existence, location and severity of a perfusion deficit. To acquire accurate results, the model has to be general enough. Hence the dataset is enlarged to 44 patients and now encompasses sufficient data to train and test on interpatient basis. To evaluate the performance of the random forest, again the intensities from 20 timepoints are employed as features. The workflow, from preprocessing to prediction, is summarized in Figure 5.1.

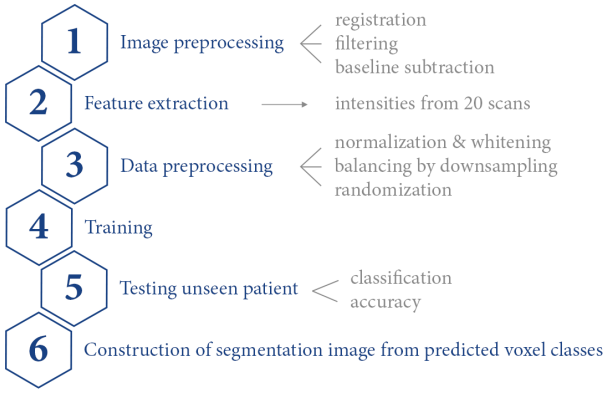



Figure 5.1: Workflow for prediction of core and penumbra using the intensities of 20 scans as features.

This is repeated 44 times, each iteration keeping a different patient aside as test set. To balance the classes, downsampling is applied. Upsampling was also tested, but gave similar results while needing much higher computation times. The amount of training samples depend on which patient is left out, but on average there are about 1 600 000 training elements per experiment.
Since the data are different from the previous section, the random forest parameters are no longer optimal. Several combinations of parameter values were tested, from which the best ones were: 100 estimators, a maximum depth of 100 and minimally 100 samples per leaf.
With these settings, the average accuracy over the leave-one-out experiment is 73.70 % with a standard deviation of 19 %. The distribution of these accuracies is illustrated by the first boxplot in Figure 5.13 on page 49. There are both good and bad results. Figure 5.2 shows a prediction with 86.92 % accuracy. The top row represents the ground truth for this patient, while the bottom row is the prediction by the random forest. A zoom on the most relevant slices is attached in Appendix B.2. The model recognizes the core and penumbra regions, but contains some misclassified voxels, spread over the brain.

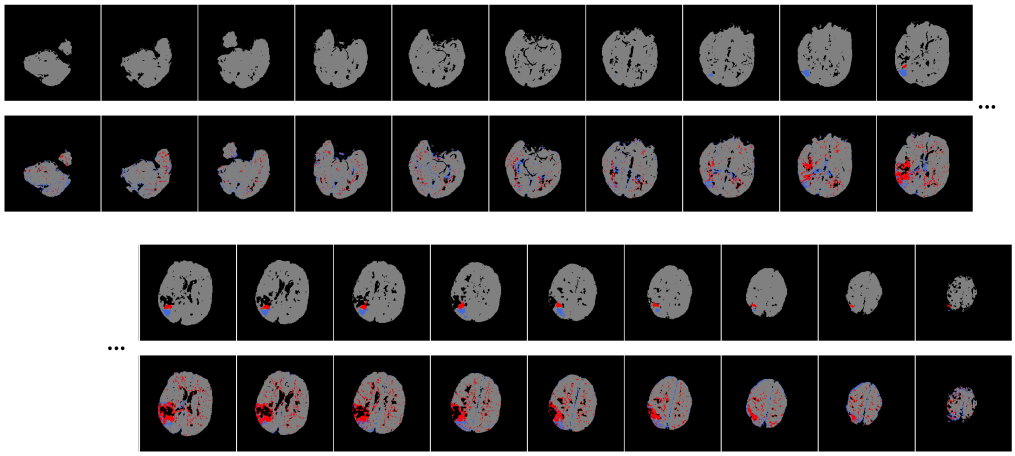


Figure 5.2: Ground truth (top row) and prediction (bottom row) with 86.92 % accuracy, resulting from a random forest using the intensities of 20 scans as features.

There are also cases for which the predictions are not correct. For example, for the patient in Figure 5.3, the accuracy is only 37.93 %. The model here classifies most of the voxels as penumbra while actually the patient is completely healthy.
In each prediction there are some misclassified voxels or small islands of voxels, scattered over the brain. In reality, a reduced perfusion area is likely to be connected through the cerebral vasculature. Other research [43, 46] where machine learning was used to automatically segment stroke lesions, on MRI or from hemorrhagic stroke, had similar observations. The predictions were postprocessed to obtain better segmentations. Here, this is accomplished by the use of morphological operations, namely erosion and dilution. Also, only the biggest connected area in 3D is kept for the core and penumbra regions. After adding this automated process to the pipeline, as illustrated in Figure 5.4, the average accuracy over the entire leave-one-out experiment is 80.61 %, with a standard

deviation of 33 %.

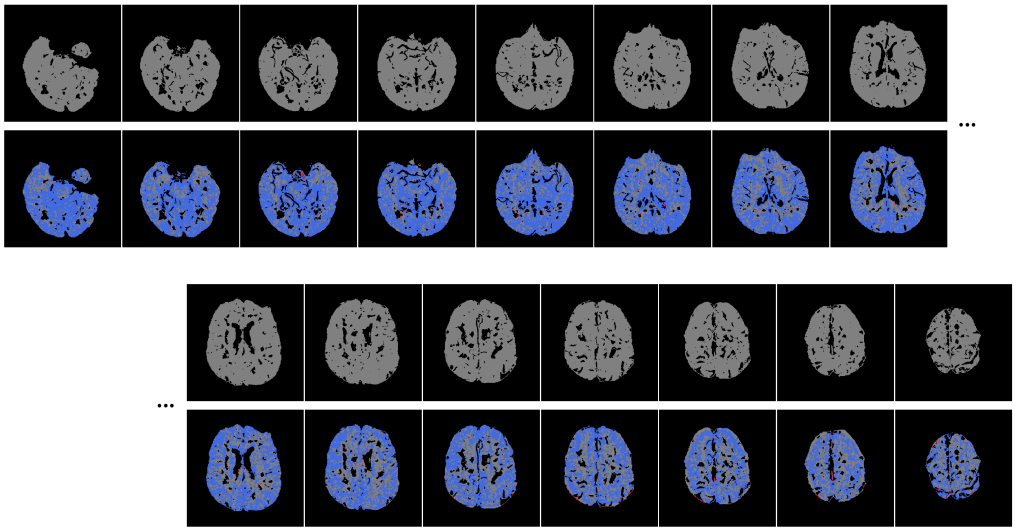


Figure 5.3: Ground truth (top row) and prediction (bottom row) with 37.93 % accuracy, resulting from a random forest using the intensities of 20 scans as features.

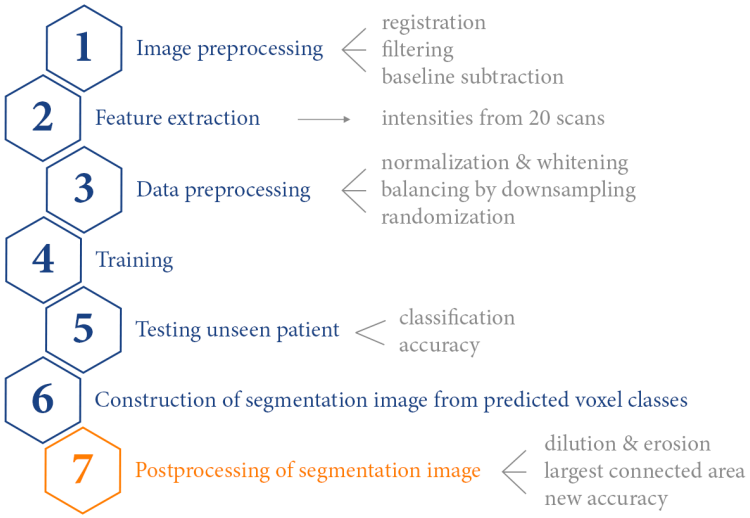



Figure 5.4: Workflow for postprocessed prediction of core and penumbra using the intensities of 20 scans as features.

The distribution is represented by the second boxplot of Figure 5.13 on page 49. The accuracy improved by 7 %, but the standard deviation increased, meaning there are greater deviations from the mean. The postprocessing enhanced the predictions that were already quite good, but deteriorated the ones that were not. This leads to a bigger gap in

accuracy between the good and bad predictions. Looking back at the previous examples, the good prediction is improved by about 12 %, now reaching 98.54 % accuracy. The bad prediction, on the other hand, is further reduced to 0 % because the morphological operations closed the areas were a few voxels were still recognized as healthy. Both are presented in Figure 5.5 and Figure 5.6 respectively. For the former, an enlarged image of the most relevant slices is added in Appendix B.2.

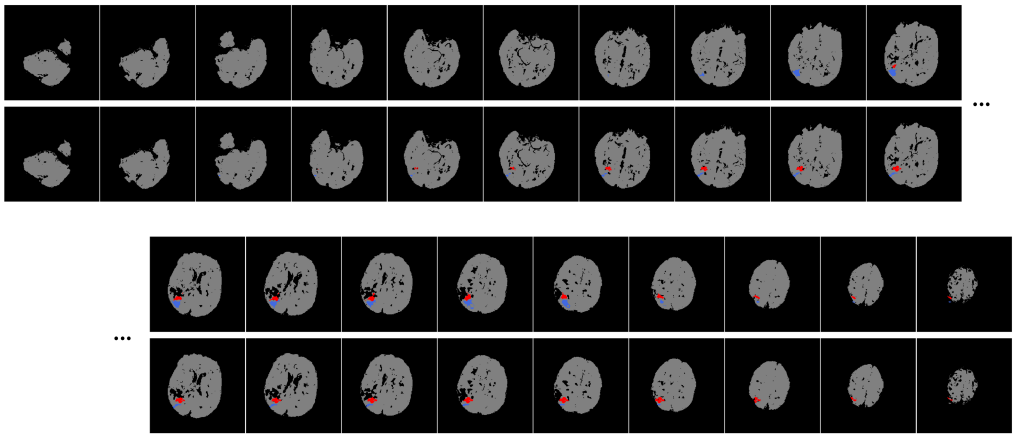


Figure 5.5: Example of the postprocessed result of a good prediction by a random forest using the intensities of 20 scans in time as features.

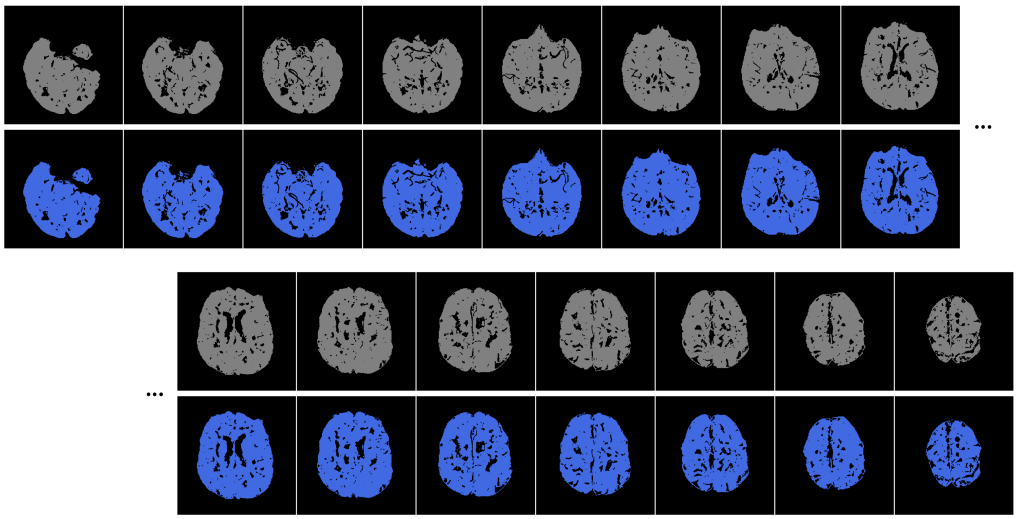


Figure 5.6: Example of the postprocessed result of a bad prediction by a random forest using the intensities of 20 scans in time as features.

The accuracy alone is not sufficient to evaluate the performance. In cases presenting a small perfusion defect, for example 10 % of the voxels, the accuracy will always be high thanks to the high amount of healthy voxels. When the model doesn't find the perfusion deficit and sees all the voxels as healthy, this will result in an accuracy of 90 %. This

seems very high, but is actually a bad result because the region of reduced perfusion is missed. To compensate for this, besides the accuracy of the classification, also the predicted core volume is considered. This metric is defined as

$$core\,volume = number\,of\,core\,voxels \quad x \quad voxel\,size \qquad [cc], \tag{5.1}$$

where the voxel size depends on the scan properties. An approximation of the size of the infarct tissue is needed as one of the parameters to base the treatment decision on. To have an idea about the acceptable error on this parameter, Dr. K. Nieboer, an emergency radiologist at the UZ Brussel, was consulted. A deviation of 10 cc from the ground truth volume is well within the limits. Applying this metric to the 44 patients, there are only three where the deviation of predicted core volume is larger than 10 cc, two of which are very close to this border. This is illustrated in Figure 5.7.

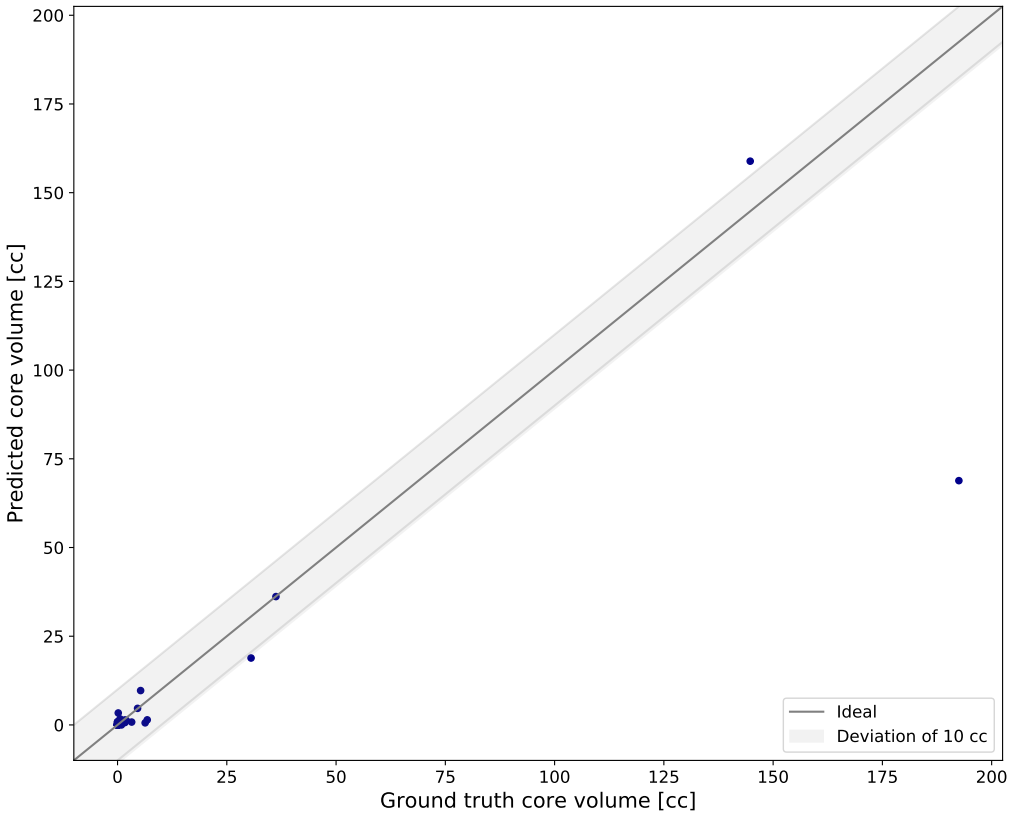



Figure 5.7: Scatter plot of ground truth core volume versus predicted core volume by a random forest using the intensities of 20 scans in time as features.

The fact that there are both good and bad predictions in this experiment, may be due to different reasons. The cerebral perfusion differs between people and it is probable that this dataset is not large enough to obtain a generalized model, containing all possible examples of each of the classes. Also, blood perfusion is a complex process. The linear decision boundaries, constructed by the random forest classifier, might not be sufficient.

More complex methods, like the ones from deep learning, may be more suited to predict this complicated process. Another option is that the features taken here, the contrast concentrations, are not sufficient. Additional features, providing more information to the machine learning algorithm, may result in better predictions.

### 5.1.2 Intensity derived features

Extra features are explored in an attempt to improve the results. First, additional features were tested that could represent the perfusion parameters. These included: the sum over all points in time, the peak value and sum of the concentrations up to a certain point in time. However, these did not improve the performance.
The research of Franzle et al. [48] concluded that the performance of a random forest classifier can be enhanced by including intensity derived features. Therefore, on top of the original contrast concentrations, filtered values are added. A kernel of size 5 x 5 x 5 voxels around the voxel of interest is considered, using four different filters: a median, a mean, a maximum and a minimum filter. For example, using the maximum filter, the value of the voxel of interest is replaced by the maximum value in its neighborhood. Integrating these four different filters will provide the random forest with additional information about the context of each sample. So, for 20 scans taken over time, this adds up to 100 features per voxel: 20 from the original intensities, plus 4 x 20 from the filtered images. Compared to the previous experiment, the workflow altered as indicated in Figure 5.8.

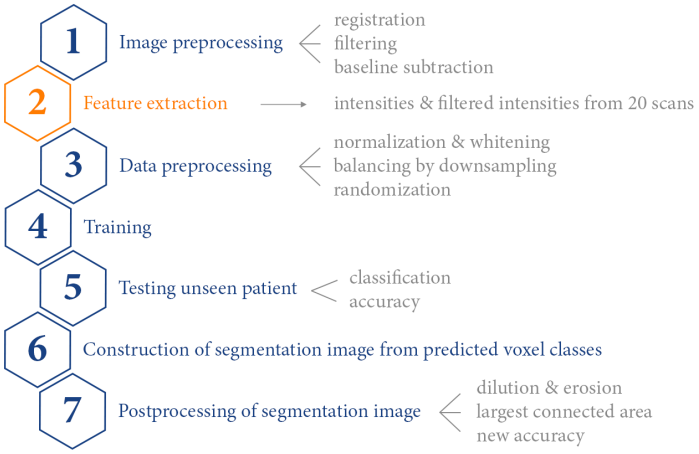



Figure 5.8: Workflow for prediction of core and penumbra using the intensities and filtered intensities of 20 scans as features.

The average accuracy over the entire leave-one-out experiment, without postprocessing, is now 84.18 % with a standard deviation of 18 %. This distribution is illustrated by the third boxplot in Figure 5.13 on page 49. Including the intensity derived features, increased the accuracy by almost 11 %. Postprocessing, however, now only raises the value to 85.32 %. This indicates that including information about the neighborhood of a voxel reduces the amount of small, scattered islands of misclassified voxels. This is also done by the postprocessing, which in this case can increase the accuracy by only 1 %,

while this was 7 % when no intensity derived features were included. This is illustrated in the following images. Figure 5.9 presents the prediction of a patient, using only the intensities as features, while Figure 5.10 shows the predictions for the same patient, including also intensity derived features. There's a clear reduction of the amount of scattering of misclassified voxels.

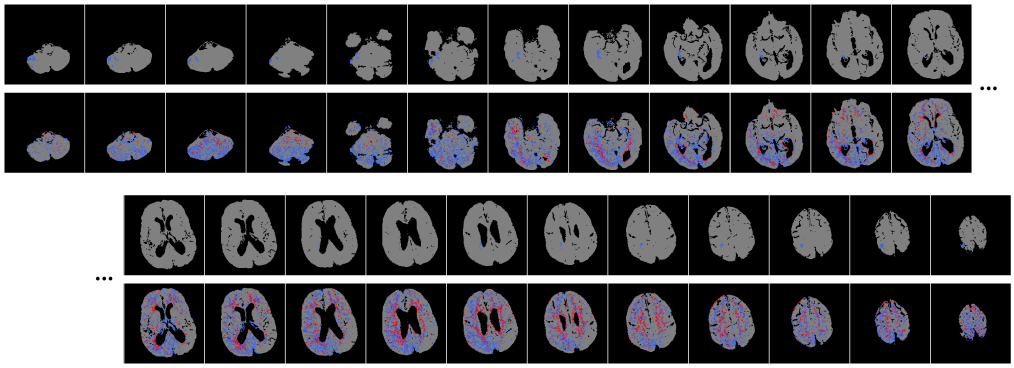


Figure 5.9: Ground truth (top row) and prediction (bottom row) with 72.05 % accuracy, resulting from a random forest using the intensities of 20 scans as features.

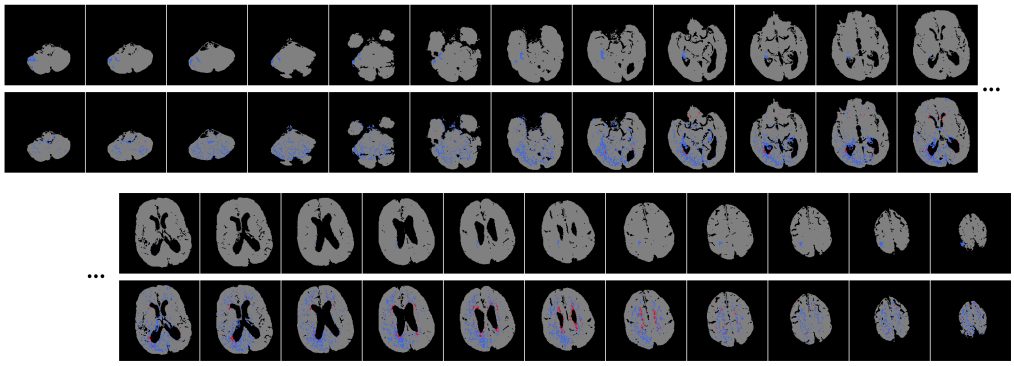


Figure 5.10: Ground truth (top row) and prediction (bottom row) with 88.46 % accuracy, resulting from a random forest using the intensities of 20 scans, together with intensity derived features.

There is, however, also a downside to adding these features. Looking at the predicted core volumes, there are now 8 patients where the error is larger than 10 cc, while in the previous experiment this was only 3 patients. The distribution is illustrated in Figure 5.11.

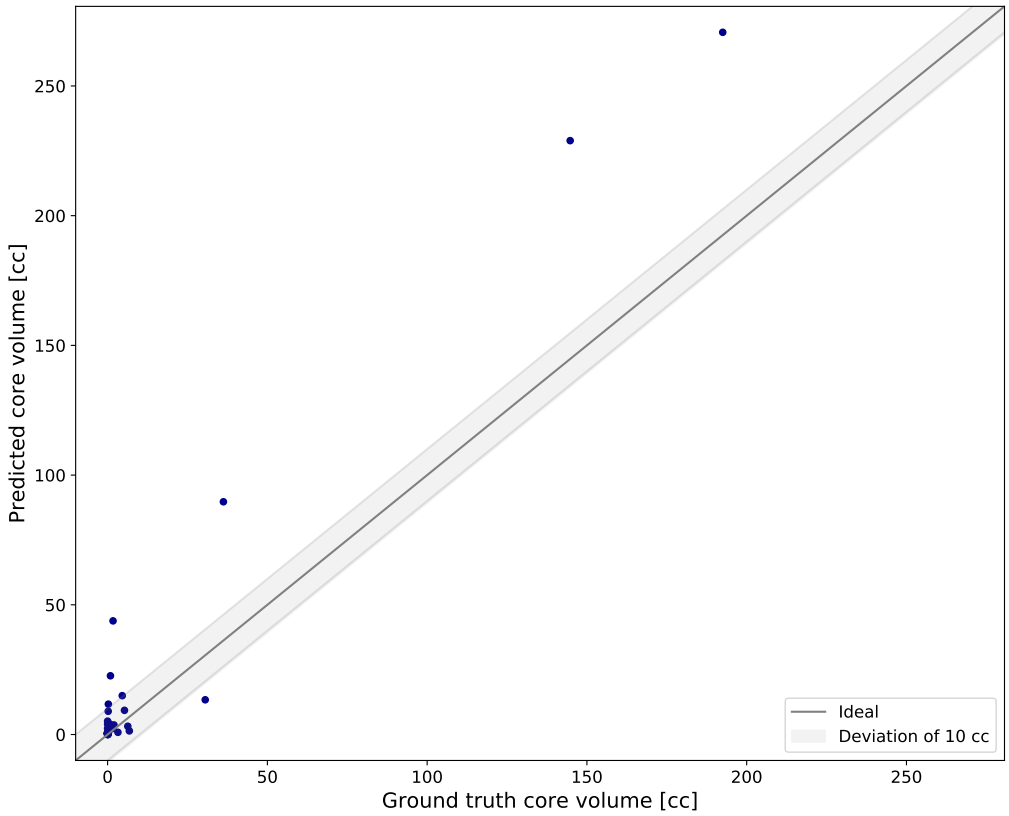



Figure 5.11: Scatter plot of ground truth core volume versus predicted core volume by a random forest using the intensities of 20 scans in time, together with intensity derived features.

The segmentations obtained by a random forest using additional intensity derived features are more interconnected, leading to more homogeneous patches of core segments. In some instances, however, the classification extends into a few slices it shouldn't. In the situation where only the intensities were used as features, some of these extension were separate patches that were removed by the postprocessing. Due to the connections in this case, these are kept for the final result, leading to some more overestimations of volumes. To illustrate this, a few slices of one of the outliers is shown in Figure 5.12. The first prediction has an overestimation of the core volume of 14 cc because the segmentation extends a bit too much in the z-direction. In the second result, because of the higher connection of all the voxels, the overestimation becomes 84 cc. This while the prediction of the core itself is quite good.

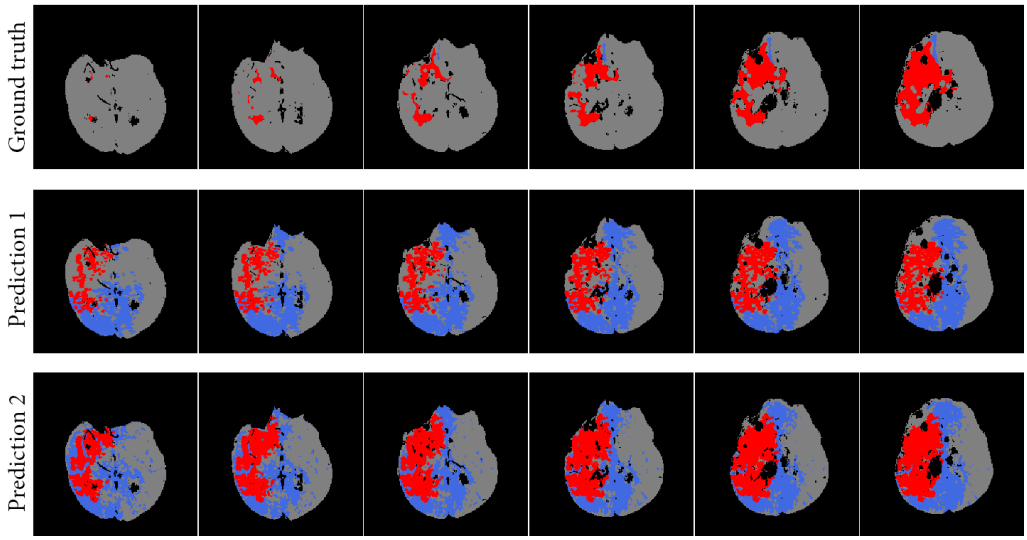



Figure 5.12: A few slices showing the ground truth (top row), prediction by random forest using the intensities as features ($2^{nd}$ row) and prediction by random forest using additional intensity derived features (bottom row).

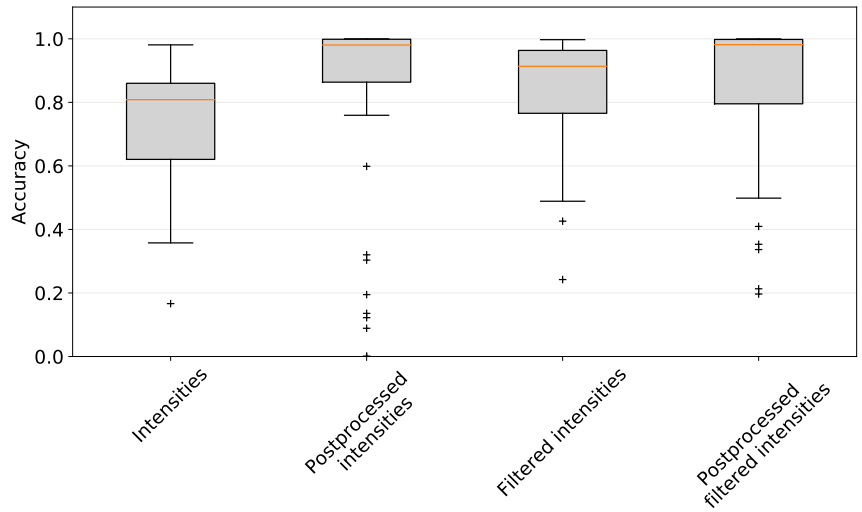



Figure 5.13: Boxplots showing the distribution of the accuracies for the leave-one-out experiments for the postprocessed segmentations of the core and penumbra.

## 5.2 Scan selection

The current perfusion CT acquired from patients suspected of having a stroke, consists of 20 scans. This is quite a lot, imposing high requirements on the scanner. If the number of scans could be reduced, the availability of the method may be increased, the radiation dose for the patient decreased and the time prior to re-perfusion lowered. In this section the influence of reducing the number of scans on the predictive power is evaluated.
Looking at a typical time-intensity curve, as the one in Figure 5.14, the first scan will not contain that much important information, as well as the last couple of scans. The number of points in time are reduced in pairs. Starting with the first and last one. Then, in every following step, the latter scans are removed. This up until the point where there are only six scans left. One, final reduction is done by removing the second and third point in time. To clarify this, Table 5.1 provides an overview of which scans are removed in which step.
As a comparison, for the case with only ten scans remaining, the impact is checked of removing every other scan as well. For clarification, the workflow is repeated in Figure 5.15. This time the 20 scans are replaced by '$x$ scans', where $x$ represents the number of scans considered. As features, only the intensities are considered since the previous experiments indicated that comparable results can be achieved with less overestimations of the core volume.

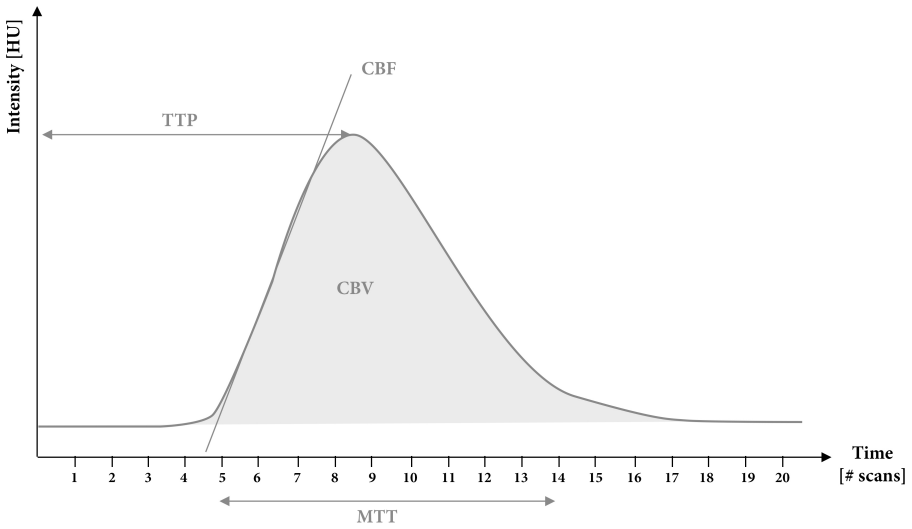



Figure 5.14: A typical perfusion graph.

| Step | Scans removed | Number of scans left |
|---|---|---|
| 1 | scans 1 & 20 | 18 |
| 2 | scans 18 & 19 | 16 |
| 3 | scans 16 & 17 | 14 |
| 4 | scans 14 & 15 | 12 |
| 5 | scans 12 & 13 | 10 |
| 6 | scans 10 & 11 | 8 |
| 7 | scans 8 & 9 | 6 |
| 8 | scans 2 & 3 | 4 |

Table 5.1: Overview of the method to remove scans.

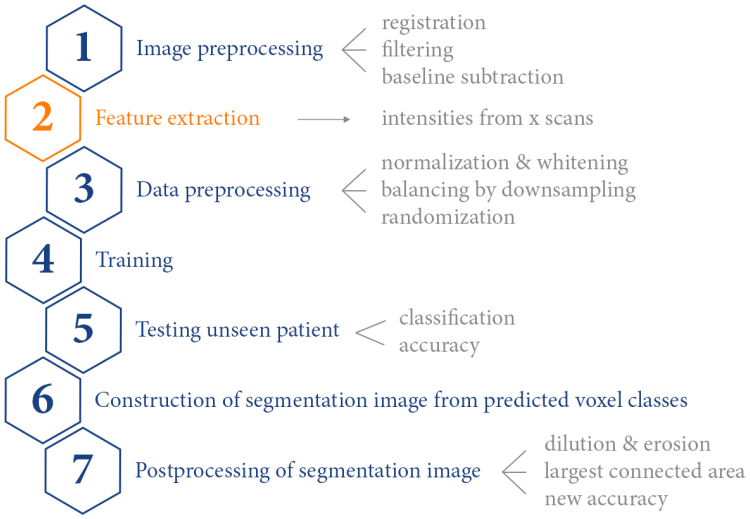



Figure 5.15: Workflow for prediction of core and penumbra using the intensities of $x$ scans as features with $x$ the number of scans considered.

Figure 5.16 illustrates the influence of removing scans on the average accuracy of the leave-one-out validation. The prediction accuracy of the random forest decreases with the number of scans. The accuracy after postprocessing fluctuates before it starts decreasing. Ten scans still offer a similar accuracy as twenty scans. The distribution of the accuracies for each step is represented by boxplots in Figure 5.17. The deviation from the mean gradually becomes larger when reducing the number of scans. For the postprocessed cases, Figure 5.18, the distributions stay similar up until 14 scans. For 12, 10 and 8 scans the deviations are increased a bit while for 6 and 4 scans they are a lot bigger.
For the situation with 10 scans, also the approach of removing every other scan was checked, resulting in an average accuracy of 69.51 % or 82.49 % postprocessed. For the method described above, these values are: 67.06 % and 82.21 % respectively. So, reducing every other point in time offers a slightly better result.

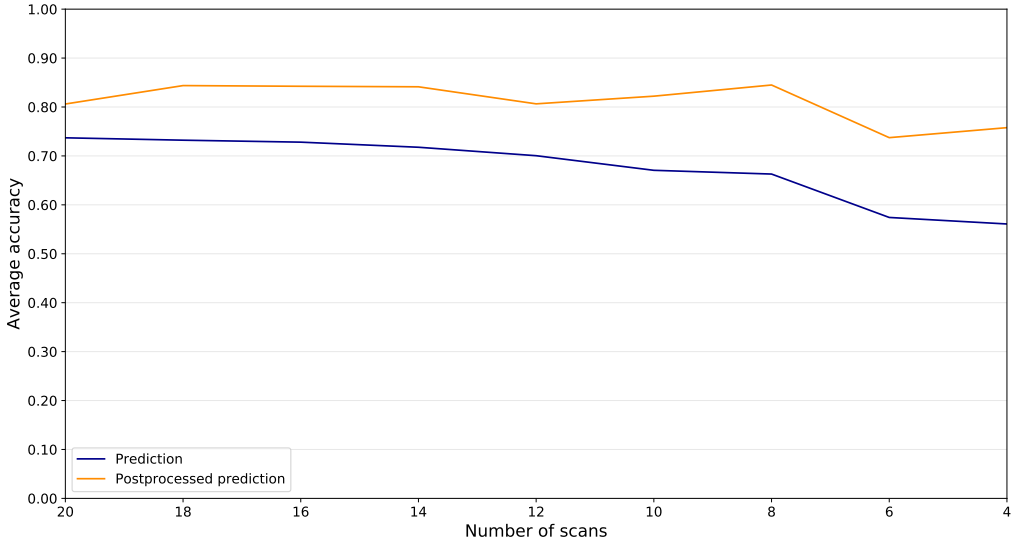



Figure 5.16: Influence of removing scans on the average accuracy of a random forest predicting three classes.

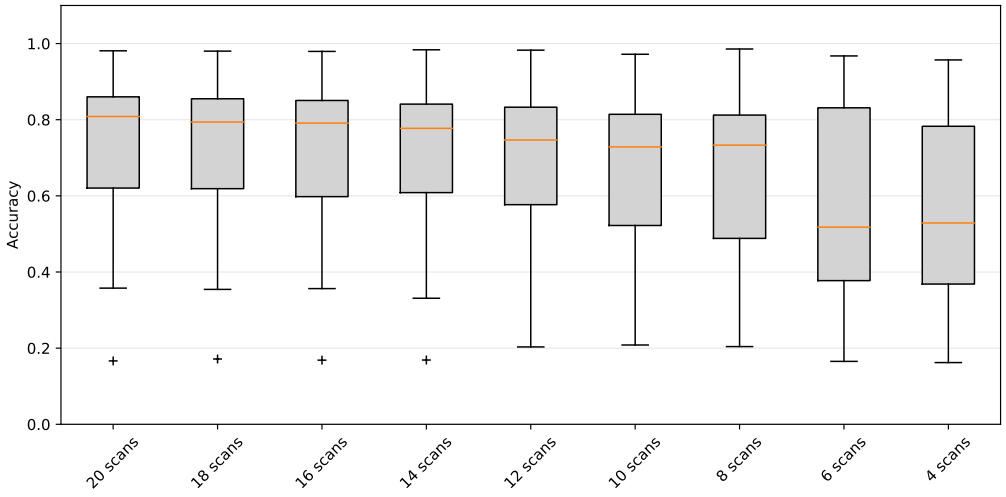



Figure 5.17: Boxplots showing the distribution of the accuracies for the leave-one-out experiments for the segmentations of the core and penumbra per step of reducing scans.

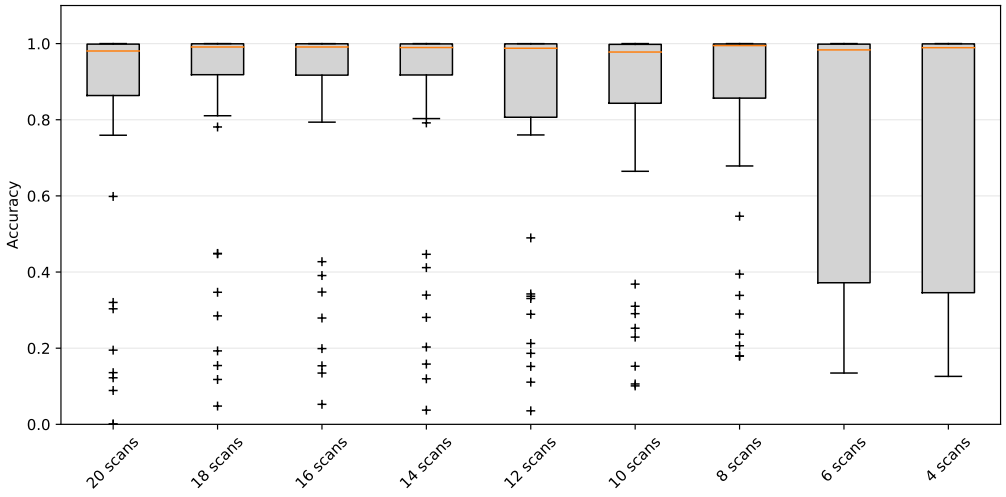



Figure 5.18: Boxplots showing the distribution of the accuracies for the leave-one-out experiments for the postprocessed segmentations of the core and penumbra per step of reducing scans.

## 5.3 Conclusion

In this chapter the dataset was expanded to 44 patients and the classification of the three classes was analyzed. A random forest using the baseline subtracted intensities of 20 scans as features already performed quite well. The leave-one-out validation over all patients obtained an average accuracy of 73.70% or even 80.61% after postprocessing. This proves the machine learning algorithm is capable of learning from perfusion CT data and can apply this knowledge to a new patient. There were however some bad results as well.
The addition of filtered intensities, including information of a neighborhood of size 5 x 5 x 5 for each voxel, increased the prediction accuracy of the random forest to 84.18 %. Postprocessing now did not have such a high impact, leading to an average accuracy of 85.32 %. Both are still higher than the results when only the intensities are used. However, looking at the prediction of the core volume, the use of only the intensities seems to be better. In this case there are only three patients where the deviation from the ground truth is larger than 10 cc. When adding the intensity derived features, this number goes up to eight. This can be explained by considering the behavior of both classifiers. When a segmentation spreads into some additional slices, the deviation of core volume quickly increases. The additional features result in more homogeneous, connected lesion areas that are not affected that much by the postprocessing. In case of only taking intensities as features, there are more separate patches that are removed in the postprocessing, resulting in less overestimations.
Reducing the number of time points indicated that there is potential to have less than twenty scans. When postprocessed, the average accuracy for ten scans is similar to the one for twenty. Removing every other scan was observed to be slightly better than the approach described above.
The fact that both good and bad predictions are present, suggests that the dataset is not diverse enough to obtain a general model. Cerebral perfusion is a very complex process. Everyone has a slightly different blood flow and a deficit can occur anywhere and have any shape or size. Probably, adding more patients to the dataset, with a variety of lesions, will yield a better performance. A second possible method to increase the accuracies is to use a more complex classifier. A deep learning algorithm can construct more complicated decision boundaries and might be more suited for these data.

# Chapter 6

# Predicting the therapeutic decision

An important parameter for the therapeutic decision is the volume of the core lesion. Recent clinical trials [49, 50] have indicated the benefit of a thrombectomy combined with standard intra-arterial therapy when the patient meets certain requirements. One of these requirements is a core volume smaller than 70 cc. Although also the size of the penumbra is a contributing factor, here the focus will be on the volume of the infarct zone. In this chapter, the performance of binary classifications is evaluated, aiming at a correct prediction of the optimal therapeutic decision based on the core volume. Both sets of features, intensities with or without intensity derived features, are analyzed, as well as the reduction of the amount of scans.

## 6.1 Binary classification

The binary classification is analyzed for each of the three classes, using the intensities as features. The core is the focus of this chapter, though also the binary classification of penumbra and healthy are checked for completeness. Possibly a three class prediction split up in binary problems can achieve better results than segmenting three classes at once.

### 6.1.1 Core prediction using intensities as features

From the 44 patients, there are sixteen for whom the reduced blood perfusion resulted in a core. Performances were compared in case the entire dataset is employed and in case it is reduced to the patients presenting a core. When all data were used, the accuracies were 2 to 5 % higher. This was due to the fact that the model has a good negative predictive value and the majority of the completely healthy patients achieved accuracies of 99 to 100 %. For the results to be more representative for the prediction of core, the dataset is reduced to those sixteen patients.
To examine the difference in performance between a three- and two-class problem, the same steps are taken as in the previous chapter. Initially a random forest is trained only on the intensities of the images from twenty scans, according to the workflow in Figure 6.1.

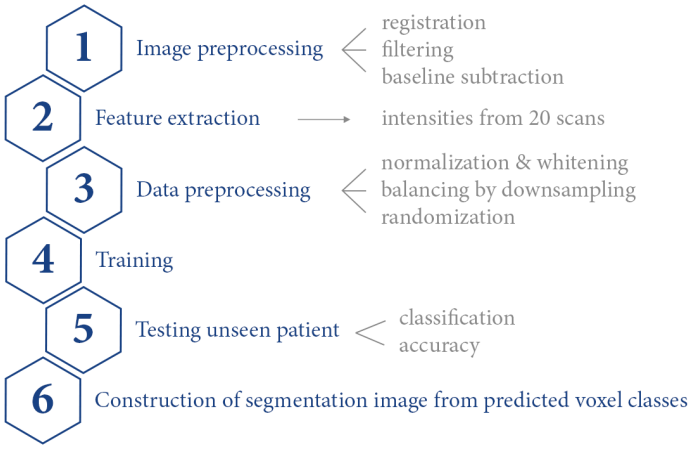



Figure 6.1: Workflow for prediction of core using the intensities of 20 scans as features.

The average accuracy obtained by a leave-one-out validation is 83.97 %, with a standard deviation of 6.8 %. The distribution is represented by the first boxplot in Figure 6.18 on page 67. Compared to the three class segmentation, the accuracy is higher and the standard deviation lower. An example of a good prediction, with 91.29% accuracy is presented in Figure 6.2. An enlarged image of the most relevant slices is added in Appendix C.2. The model clearly recognizes the core region, but again some small patches of misclassified voxels are present. The lowest accuracy achieved in this experiment is 71.45%, for the patient in Figure 6.3. There are a lot of voxels considered core that are actually healthy. However, the correct half does have a higher concentration of core voxels.

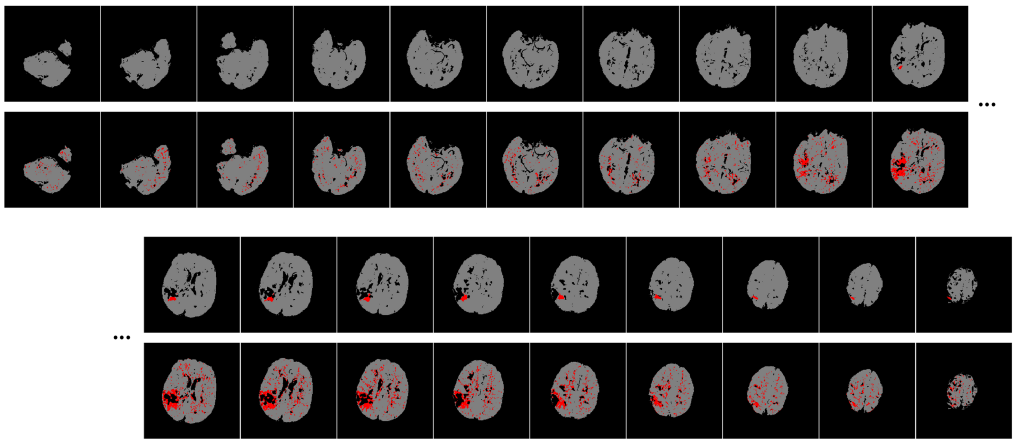


Figure 6.2: Ground truth (top row) and prediction (bottom row) of a core segmentation with 91.29% accuracy, resulting from a random forest using the intensities of 20 scans as features.

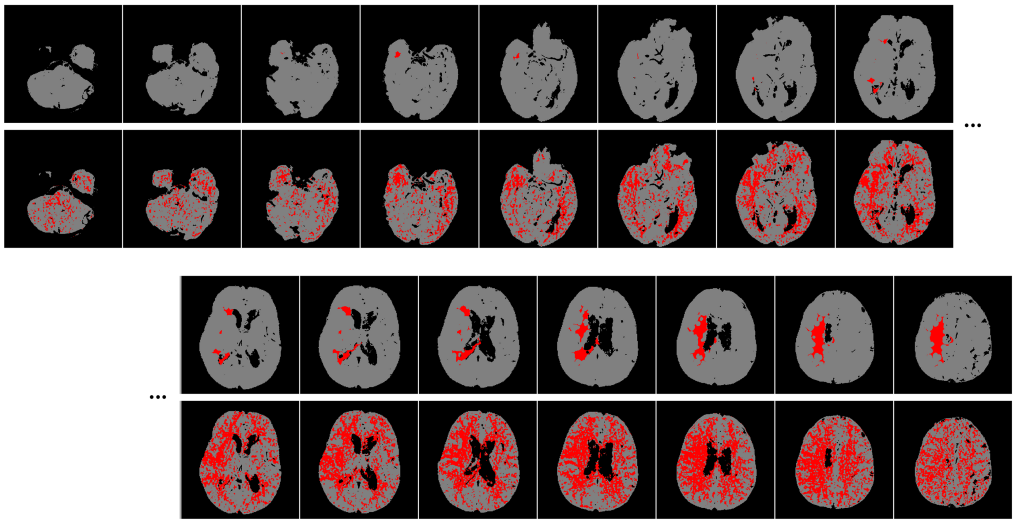


Figure 6.3: Ground truth (top row) and prediction (bottom row) of a core segmentation with 71.45% accuracy, resulting from a random forest using the intensities of 20 scans as features.

The model seems to recognize the problem area, but possibly this patient has an overall reduced cerebral blood flow compared to the other patients in the dataset. Therefore, the model wrongly sees some of the healthy voxels as core. Again, the question rises if the dataset is general enough to obtain higher results for all patients.
To remove these islands of core, postprocessing is added to the pipeline, as indicated in Figure 6.4.

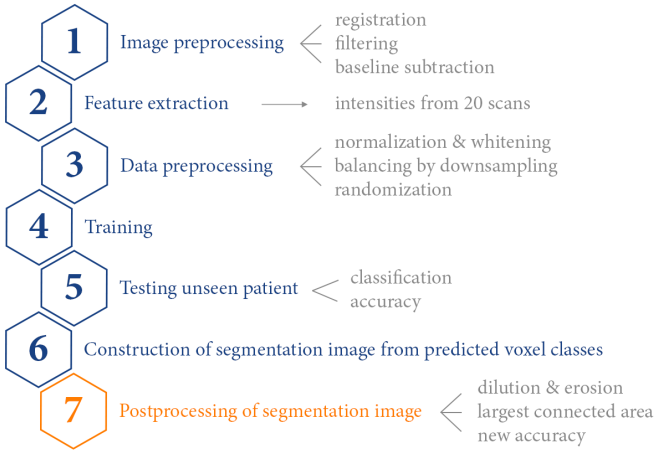



Figure 6.4: Workflow for postprocessed prediction of core using the intensities of 20 scans as features.

This step increases the average accuracy to 96.82 %, with a standard deviation of 5.2 %. The distribution is illustrated by the second boxplot in Figure 6.18 on page 67. The same patients are considered again in Figure 6.5 and Figure 6.6 respectively. For the former,

a zoom on the most relevant slices is added in Appendix C.2. Both segmentations now have a much better delineation. The first patient has a nicely defined core, without the scattered patches, leading to an accuracy of 99.19 %. The second patient also has a much clearer segmentation of the core, increasing the accuracy from 71.45% to 93.80%. This proves that the model did recognize the location of the core, while a lot of the misclassified voxels were not interconnected.

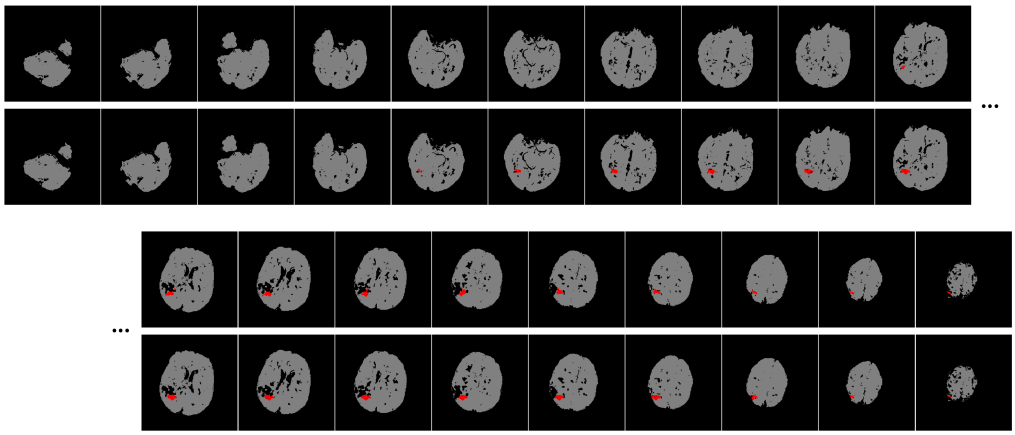


Figure 6.5: Ground truth (top row) and postprocessed prediction (bottom row) of a core segmentation with 99.19% accuracy, resulting from a random forest using the intensities of 20 scans as features.

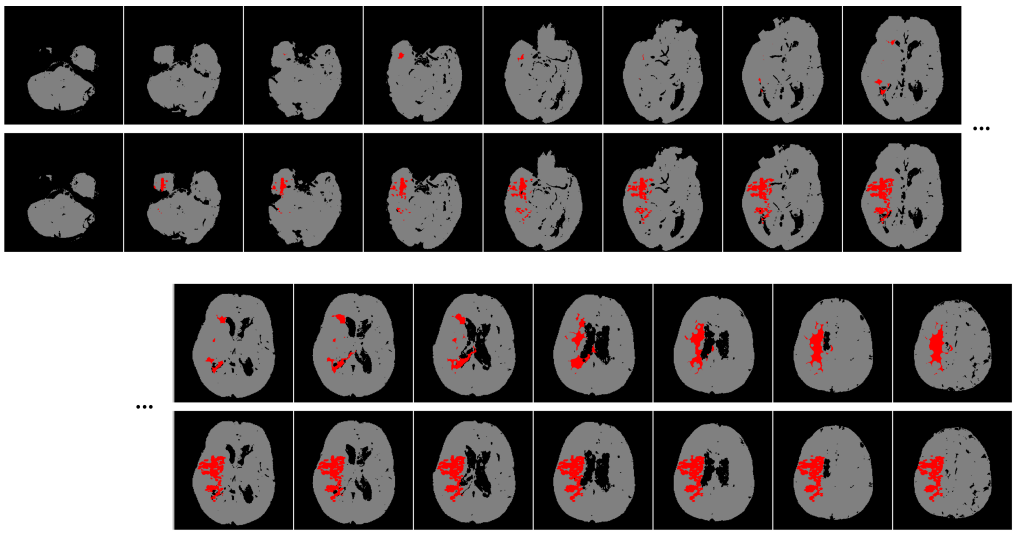


Figure 6.6: Ground truth (top row) and postprocessed prediction (bottom row) of a core segmentation with 93.80% accuracy, resulting from a random forest using the intensities of 20 scans as features.

In the previous chapter, where the three class problem was tackled, there were still some bad classifications with very low accuracies. Now, only considering two classes, not only

the average accuracy is much better, but also the lowest accuracy is 83.20%.
Important to note, however, is that there is one patient where the segmentation is not correct, despite a high accuracy. This is shown in Figure 6.7. The size is comparable to the actual core and the accuracy is 99.69 % thanks to the high amount of healthy voxels. However, the location of the lesion is wrong. This is the only case, out of the dataset of sixteen patients, where this occurs.

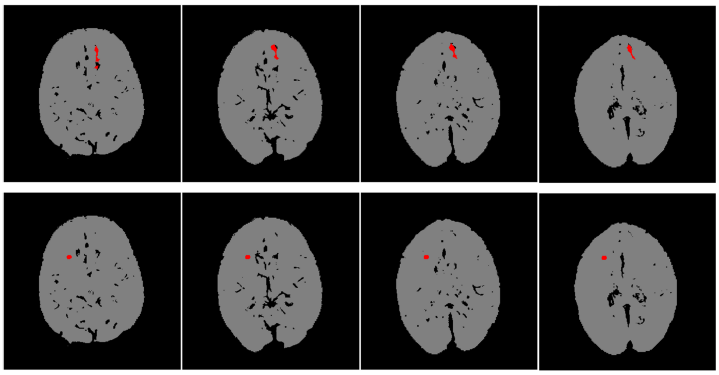


Figure 6.7: A few slices showing the ground truth (top row) and postprocessed prediction (bottom row) with high accuracy and comparable lesion size, but wrong location.

This error may be due to the small training dataset. Sixteen patients might not be enough for a general model. Also, the core volume is very small, so there is not a clear, big region with a perfusion deficit. However the model should be able to find these subtle differences in blood flow, as it does for other patients in the dataset, for example the one in Figure 6.8.

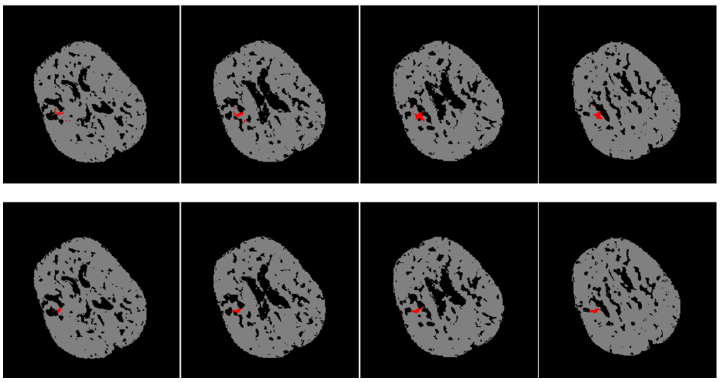


Figure 6.8: A few slices showing the ground truth (top row) and postprocessed prediction (bottom row) of a core segmentation with 99.79 % accuracy, resulting from a random forest using the intensities of 20 scans as features.

This segmentation with a high accuracy, but wrong location confirms that the accuracy alone is not sufficient to evaluate the predictions. As in the previous chapter, the prediction of the core volumes is assessed. A scatter plot of the ground truth versus predicted core volume is presented in Figure 6.9. The threshold of 70 cc is indicated with the red lines. There are four patients where the deviation of the volume is larger than 10 cc, each time overestimating the size. There are only two patients where the prediction of volume smaller or bigger than 70 cc is not correct.

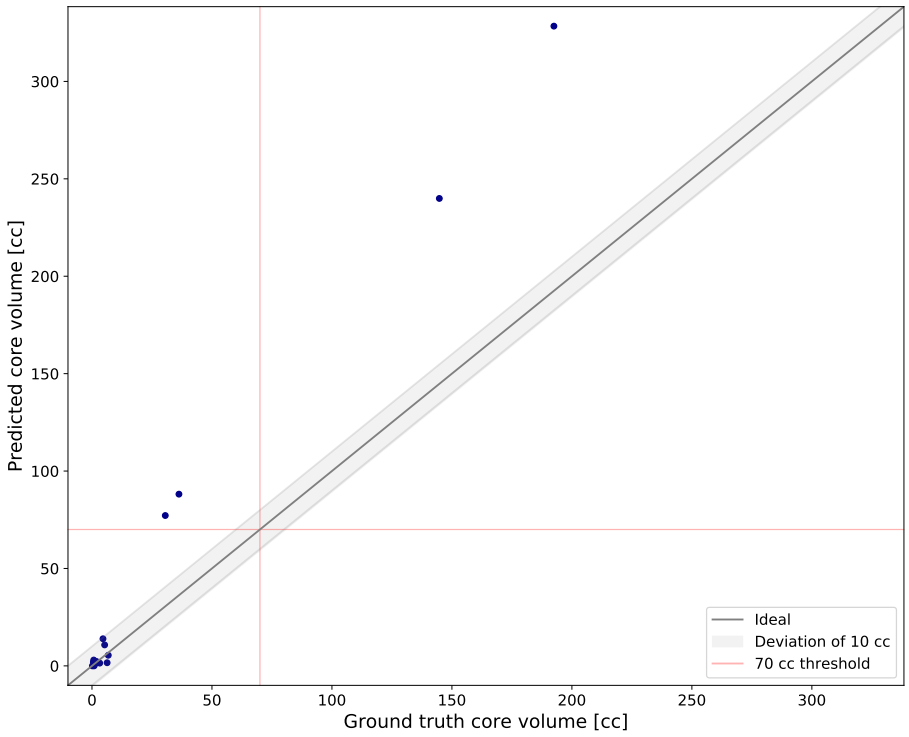



Figure 6.9: Scatter plot of ground truth versus predicted core volume for core segmentations by a random forest using the intensities of 20 scans in time as features.

The four cases with a deviation of the core volume larger than 10 cc are not necessarily bad predictions. For example, the patient with the biggest deviation is shown in Figure 6.10. The overestimation is solely due to the fact that the predicted core extends in a few more slices. With large lesions, any extra slice where the core is extended, leads to a huge increase in deviation. This example depicts the prediction of a random forest using only the intensities as features, no intensity derived features, and before postprocessing. Therefore, this overestimation cannot be accounted to any additional information on the neighborhood of the voxels. Probably the transition between healthy and core tissue is not clearly delineated and the random forest misclassified some of these transition voxels.

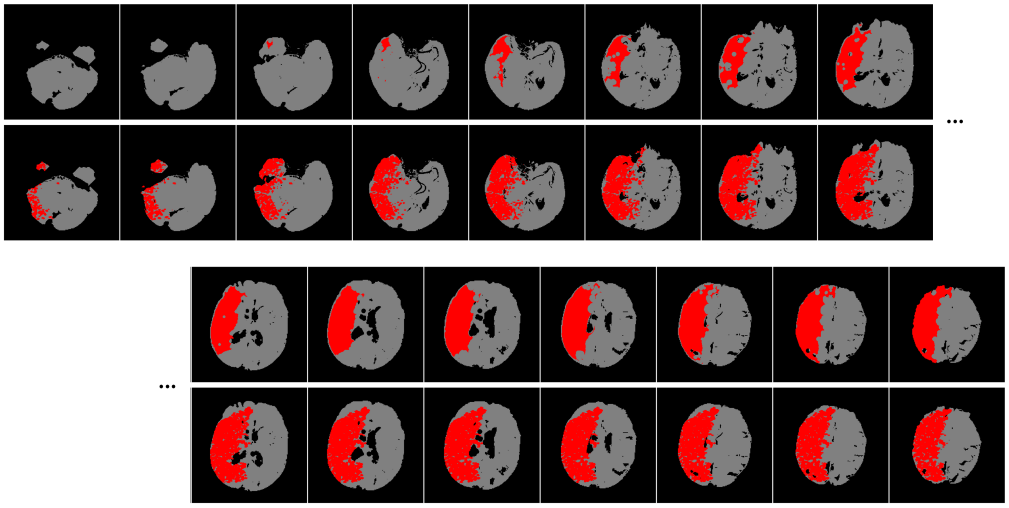


Figure 6.10: Ground truth (top row) and postprocessed prediction (bottom row) of a core segmentation with a big lesion size.

This assessment of the core volume already provides extra information on the quality of the classification. Yet, knowledge on the correctness of the predicted location is lacking. Other metrics that are often used in segmentation evaluations are: sensitivity, specificity and dice score. These are defined as:

$$sensitivity = \frac{TP}{TP + FN}, \tag{6.1}$$

$$specificity = \frac{TN}{TN + FP}, \tag{6.2}$$

$$dice = \frac{2 * TP}{2 * TP + FP + FN}, \tag{6.3}$$

with $TP$ the true positives, $TN$ the true negatives, $FP$ the false positives and $FN$ the false negatives. Applied to these data, the average sensitivity is only 0.30, the specificity is 0.97 and the dice is 0.22. The specificity is high thanks the high amount of negative voxels, validating the ability of the model to recognize healthy voxels. The sensitivity and dice score, on the other hand, are low. However, looking at Figure 6.11, a case with a sensitivity of 0.18 and a dice score of 0.17, the prediction of the core lesion is actually quite accurate in terms of size and location and of use for clinical decision making. Due to the small size and limited overlap, metrics that emphasize true positive prediction yield very low values.

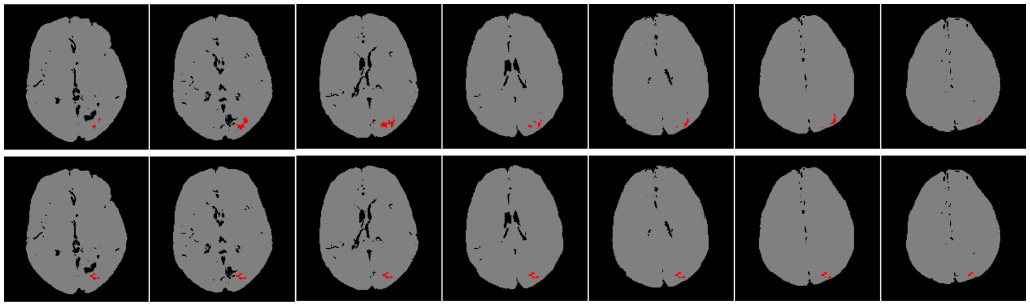


Figure 6.11: A few slices showing the ground truth (top row) and postprocessed prediction (bottom row) of a good core segmentation with a low sensitivity and dice score.

For a proper evaluation of the segmentation, a different metric is needed. To evaluate the precision of the predicted location, the center of mass is determined for both the ground truth and the classification. The *center_of_mass* function from the *ndimage.measurements* package of *Scipy* returns the (x,y,z)-coordinates of this central point. The localization error is then defined as the distance in millimeters between both. Regarding the volume, the relative error is defined as

$$relative\,volume\,error\;=\;\frac{V_{predicted}\;-\;V_{ground\,truth}}{average(V_{predicted},\,V_{ground\,truth})}, \tag{6.4}$$

with $V$ the volume. This prevents small deviations to have a huge impact on the error, solely because of the lesion size.
The scatter plot in Figure 6.12 illustrates the results for these metrics with on each axis a gray line indicating their mean value. The biggest localization error is indeed made by the prediction that was shown in Figure 6.7. The majority of the predictions are very well localized, not exceeding an error of 2 cm.

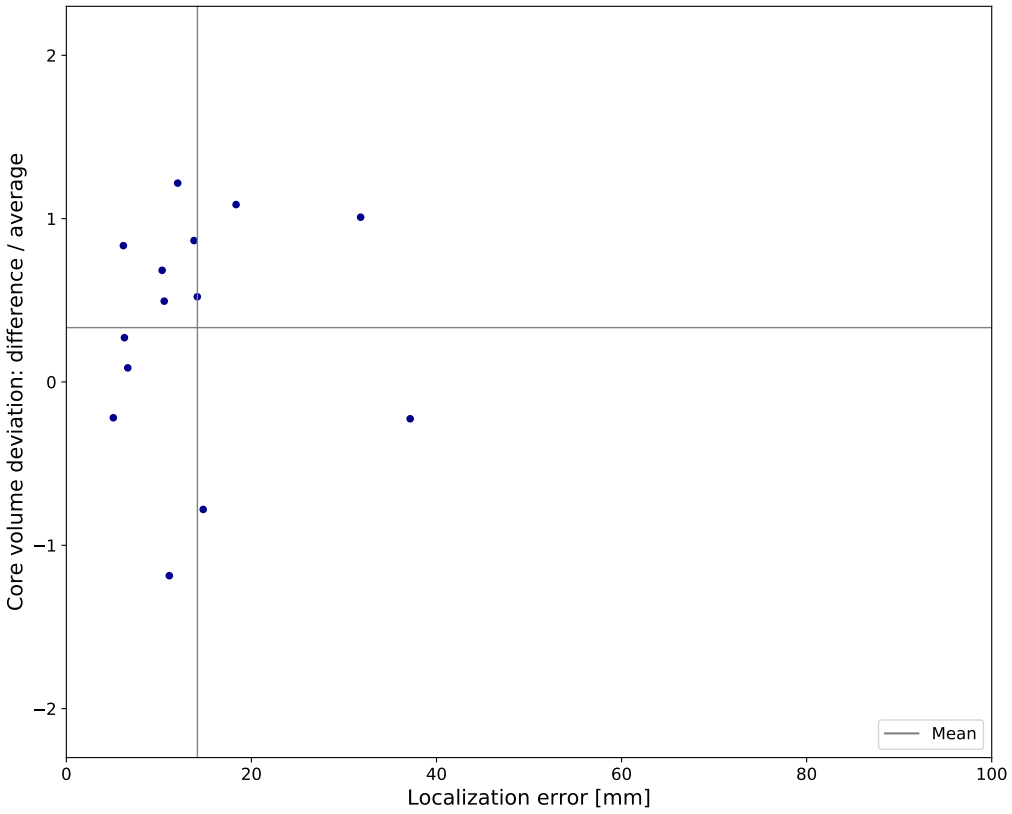



Figure 6.12: Scatter plot of localization error versus the relative error in core volume estimation for core segmentations by a random forest using the intensities of 20 scans in time as features.

### 6.1.2 Penumbra prediction using intensities as features

The binary classification is also tested to predict the penumbra. In order not to have an increase in accuracy from the completely healthy patients, the dataset is reduced to the cases presenting a penumbra region. Now, the dataset comprises 24 patients and the leave-one-out experiment yields an average accuracy of 76.94 %. This is only slightly better than the average accuracy of the same experiment with three classes, 73.70 %. For some patients, the segmentation is very accurate. However, for others, again the entire brain is considered to be penumbra. Because of this, postprocessing cannot increase the average accuracy any further. Only for certain patients it results in a good prediction of the penumbra, like the one in Figure 6.13. So for penumbra, this method does not offer an improvement from the three class results.

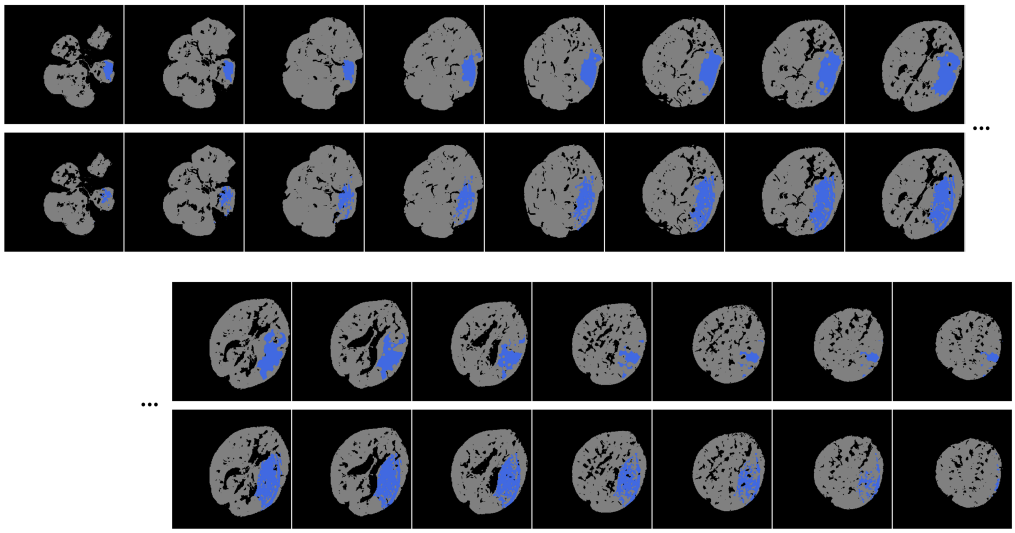


Figure 6.13: Ground truth (top row) and prediction (bottom row) of a penumbra segmentation with 91.90% accuracy, resulting from a random forest using the intensities of 20 scans as features.

### 6.1.3 Healthy prediction using intensities as features

Prediction of healthy tissue is just as important as predicting an infarction or reduced perfusion area. Hence, an analysis of the binary classification of healthy or not healthy is performed as well. For this, the dataset consists of the 44 patients again. The average accuracy of the leave-one-out experiment yields 81.43 %. This is an improvement from the 73.70 % of the three class segmentation. Postprocessing increases the performance further to 87.25 %. Some patients can achieve 90 % accuracy, even before postprocessing, like the one in Figure 6.14. Some other patients, however, perform very bad and are seen as completely unhealthy, like in Figure 6.15.

### 6.1.4 Conclusion

Although the average accuracy improved by splitting the three class problem into three binary classifications, for penumbra and healthy, the dataset seems to be too small. There are some patients with good results and some with bad ones for these two classes. More examples of the various possible perfusion deficits are required to obtain a more general model. The binary classification of core or no core performed best and will be investigated further.

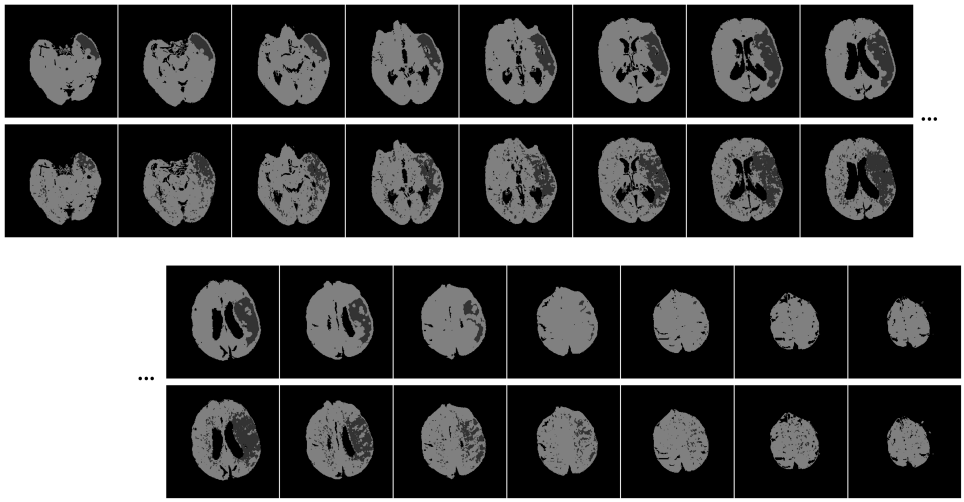


Figure 6.14: Ground truth (top row) and prediction (bottom row) of a healthy or not healthy segmentation with 90.28% accuracy, resulting from a random forest using the intensities of 20 scans as features.

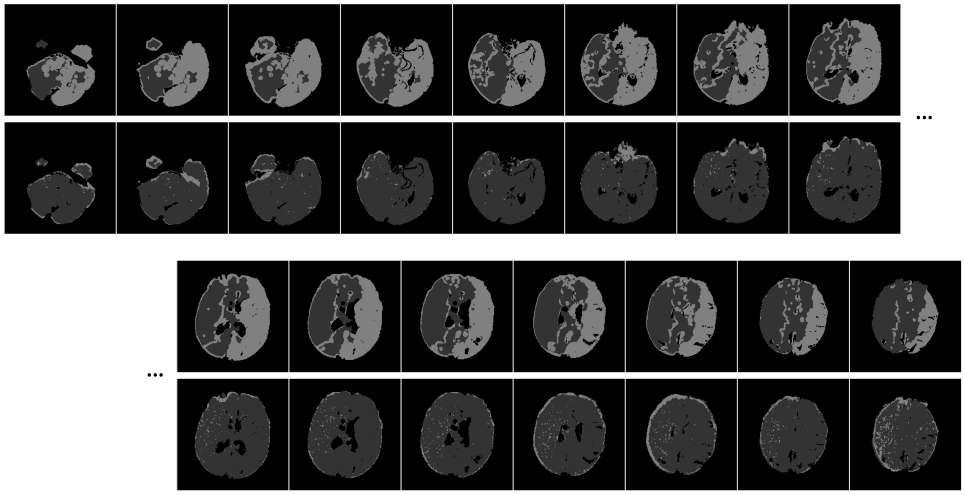


Figure 6.15: Ground truth (top row) and prediction (bottom row) of a healthy or not healthy segmentation with 47.18% accuracy, resulting from a random forest using the intensities of 20 scans as features.

## 6.2 Core prediction using additional features

Aiming to improve the results, the additional intensity derived features are provided to the random forest and the workflow from Figure 6.16 is followed.

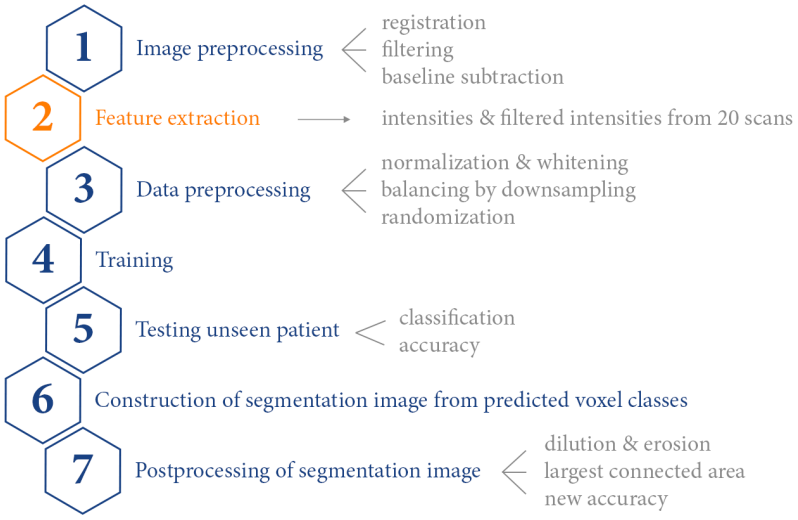



Figure 6.16: Workflow for prediction of core using the intensities and filtered intensities of 20 scans as features.

The average accuracy for the leave-one-out validation, before postprocessing, is increased from 83.97 % to 93.06 %. The distribution of the accuracies are summarized in the third boxplot of Figure 6.18. Similar to the observation in the three class segmentations, the addition of these features leads to a lower amount of scattered, misclassified voxels. For example the patient from Figure 6.2, where intensities were employed as features, is repeated in Figure 6.17, this time with the addition of the intensity derived features. An enlarged image of the most relevant slices is added in Appendix C.2.
Postprocessing of these predictions further improves the average accuracy to 96.16 %, with a standard deviation of 6 %, which is similar to the results of the postprocessed predictions using only the intensities as features. Again, this step does not have a very big impact on the average accuracy. The distribution of these accuracies is depicted by the fourth boxplot in Figure 6.18.

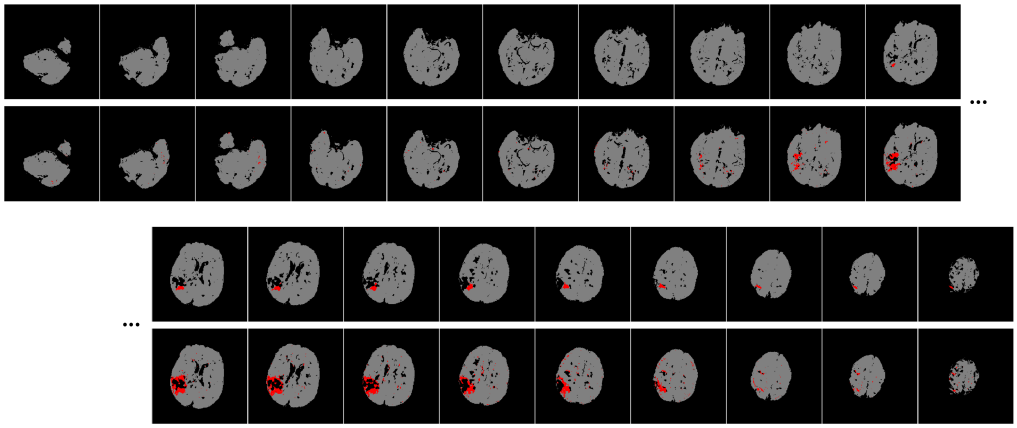


Figure 6.17: Ground truth (top row) and prediction (bottom row) of a core segmentation with 97.42% accuracy, resulting from a random forest using the intensities of 20 scans as features, together with intensity derived features.

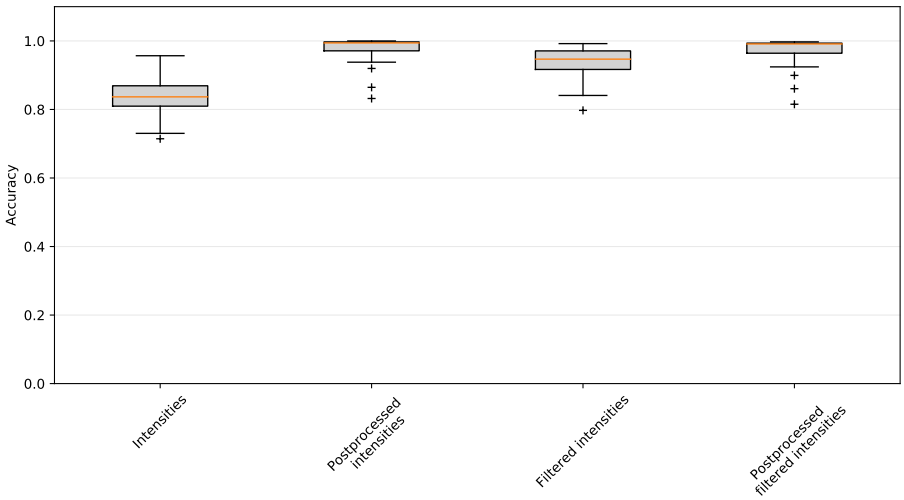



Figure 6.18: Boxplots showing the distribution of the accuracies for the leave-one-out experiments for the segmentation of the core.

Looking at the prediction of the core volume, Figure 6.19, the results are also similar. For two patients, the prediction of the treatment decision would be wrong. Any deviation of more than 10 cc from the ground truth is again due to overestimation. Now however, there are seven of these cases while previously there were only four. As seen in Chapter 5, the intensity derived features lead to more overestimations of the core volume. The same reason is valid here. Some segmentations continue into extra slices compared to the ground truth. Because of the interconnections obtained by the random forest here, the excessive volume is larger and will not be reduced in the postprocessing step. However, this does not mean the segmentation itself is bad.

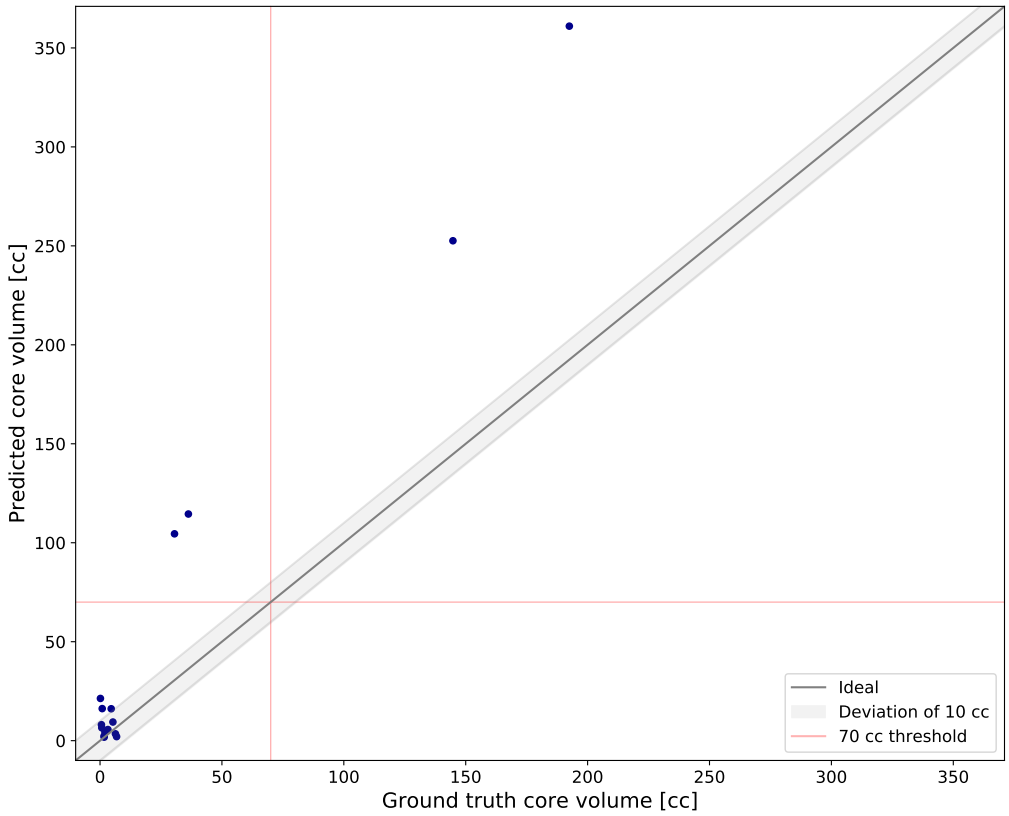



Figure 6.19: Scatter plot of ground truth core volume versus predicted core volume.

This time, for the patient from Figure 6.7 that achieved a good accuracy and predicted lesion size, but the wrong location, the correct position is found, as shown in Figure 6.20. For this patient, a bigger training set with more diverse examples of core, might offer more consistent results.

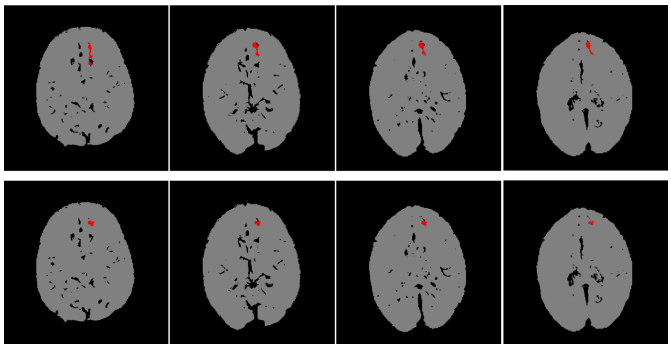


Figure 6.20: Ground truth (top row) and postprocessed prediction (bottom row) with 99.72 % accuracy, resulting from a random forest using the intensities of 20 scans as features, together with intensity derived features.

The localization error and relative difference in core volume are evaluated in Figure 6.21. None of the positioning errors exceed 3 cm.

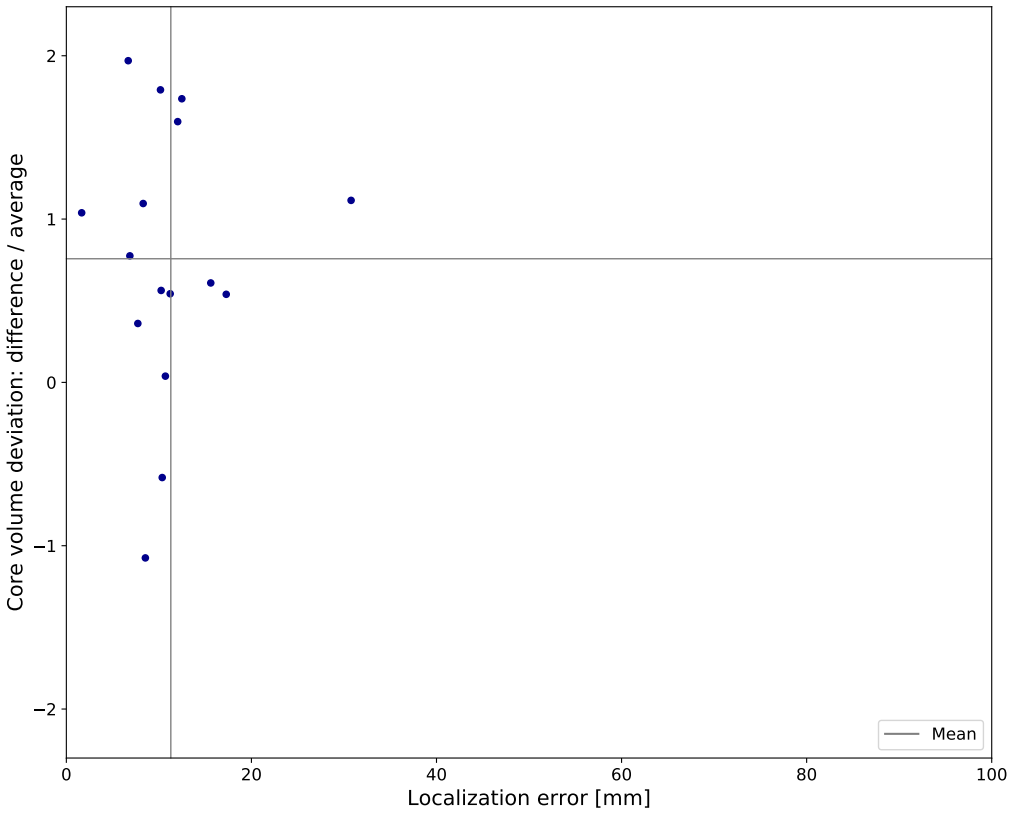



Figure 6.21: Scatter plot of localization error versus the relative error in core volume estimation for core segmentations by a random forest using the intensities of 20 scans in time as features, together with intensity derived features.

## 6.3 Core prediction: scan selection

As mentioned before, a reduction of the number of scans in the perfusion CT protocol, can increase the availability of the treatment, reduce the radiation dose for the patient and reduce the time to re-perfusion. To be able to compare the effect of removing scans to the experiments conducted in the previous sections, the same approach is adopted. The first analysis only takes into account the intensities while in a next step the filtered intensities are added.

### 6.3.1 Intensities as features

In this section multiple leave-one-out experiments are conducted according to the pipeline in Figure 6.22 with each time a different amount of scans.

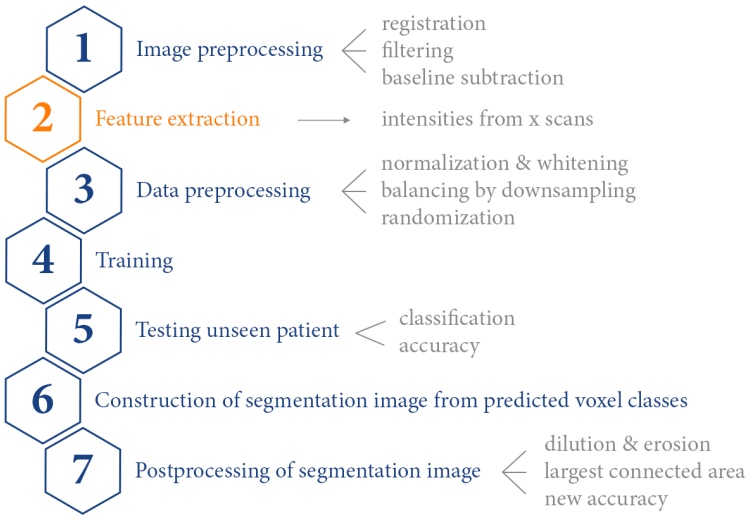



Figure 6.22: Workflow for prediction of core using the intensities of $x$ scans as features with $x$ the number of scans considered.

The impact of reducing the number of scans on the prediction by the random forest using the intensities as features, is summarized in Figure 6.23. The prediction accuracy gradually decreases with the number of scans. For the postprocessed predictions however, the accuracy fluctuates a bit, but stays above 90 %. Using ten scans still provides an average accuracy of 75.69 % and 96.04 % after postprocessing. When every other scan is removed these are: 77.76 % and 97.17 % respectively. The latter method is thus slightly better. The distribution of the accuracies is illustrated by a series of boxplots in Figure 6.24 and, for the postprocessed cases, Figure 6.25. The deviations gradually become bigger with the reduction of scans. For the postprocessed cases, the distributions are actually similar or even less dispersed up until 12 scans. With a further reduction, the deviations become slightly bigger.

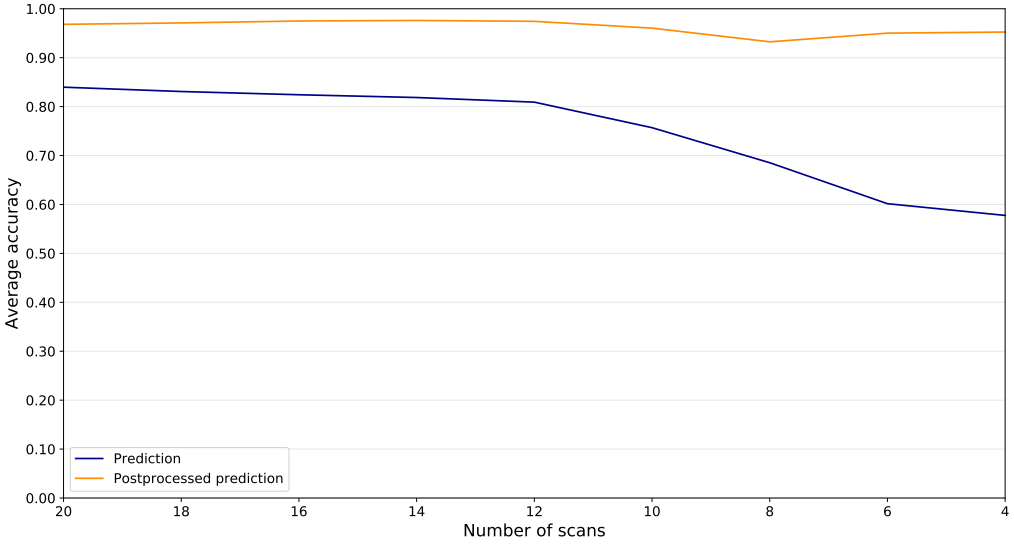



Figure 6.23: Influence of removing scans on the accuracy of a random forest using intensities as features.

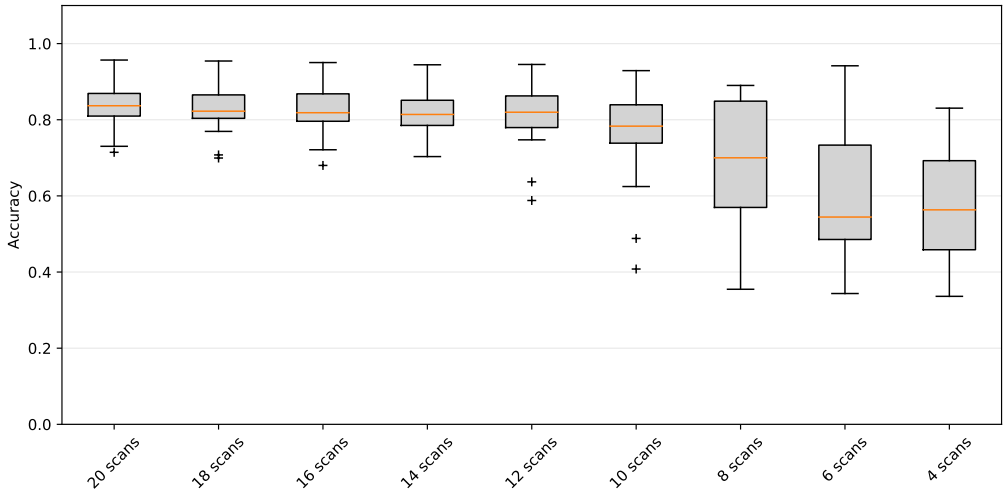



Figure 6.24: Boxplots showing the distribution of the accuracies for the leave-one-out experiments for the segmentation of the core per step of reducing scans. The features for the random forest are the intensities.

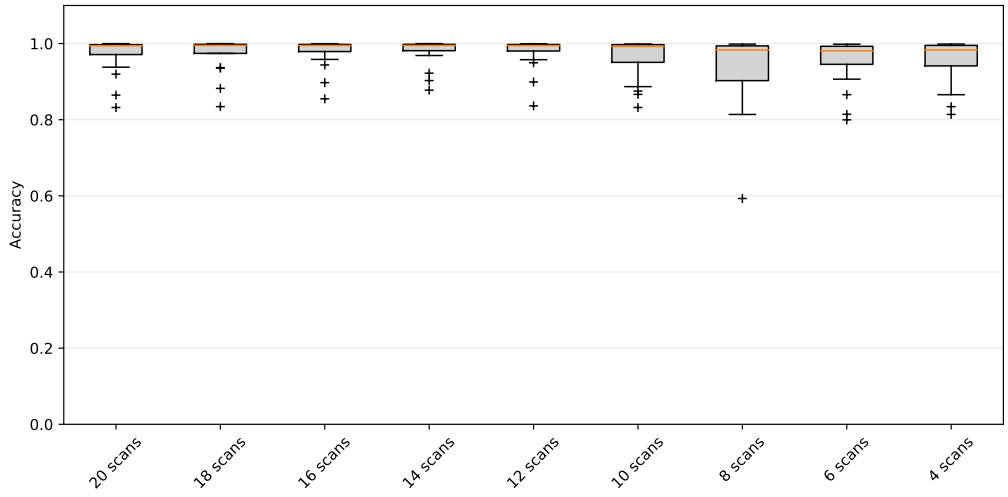



Figure 6.25: Boxplots showing the distribution of the postprocessed accuracies for the leave-one-out experiments for the segmentation of the core per step of reducing scans. The features for the random forest are the intensities.

Considering the estimation of the core volume, an overview of all steps is provided in Figure 6.26. It can be observed that reducing the number of scans actually improves the performance. With 18 scans, for 15 out of 16 patients the therapeutic decision based on the lesion volume is predicted correctly. Reducing the number of scans to 16, seems to be the most optimal because for all patients, there is a proper prediction of the therapeutic decision. Only for two patients there is an overestimation of the core volume bigger than 10 cc. The decision regarding therapy stays correct for all patients up until 12 scans. From 10 scans onwards, there's an increase in patients for which there is a wrong prediction. Also, the number of cases where the deviation from the ground truth core volume transcends 10 cc, is higher than for 20 scans.

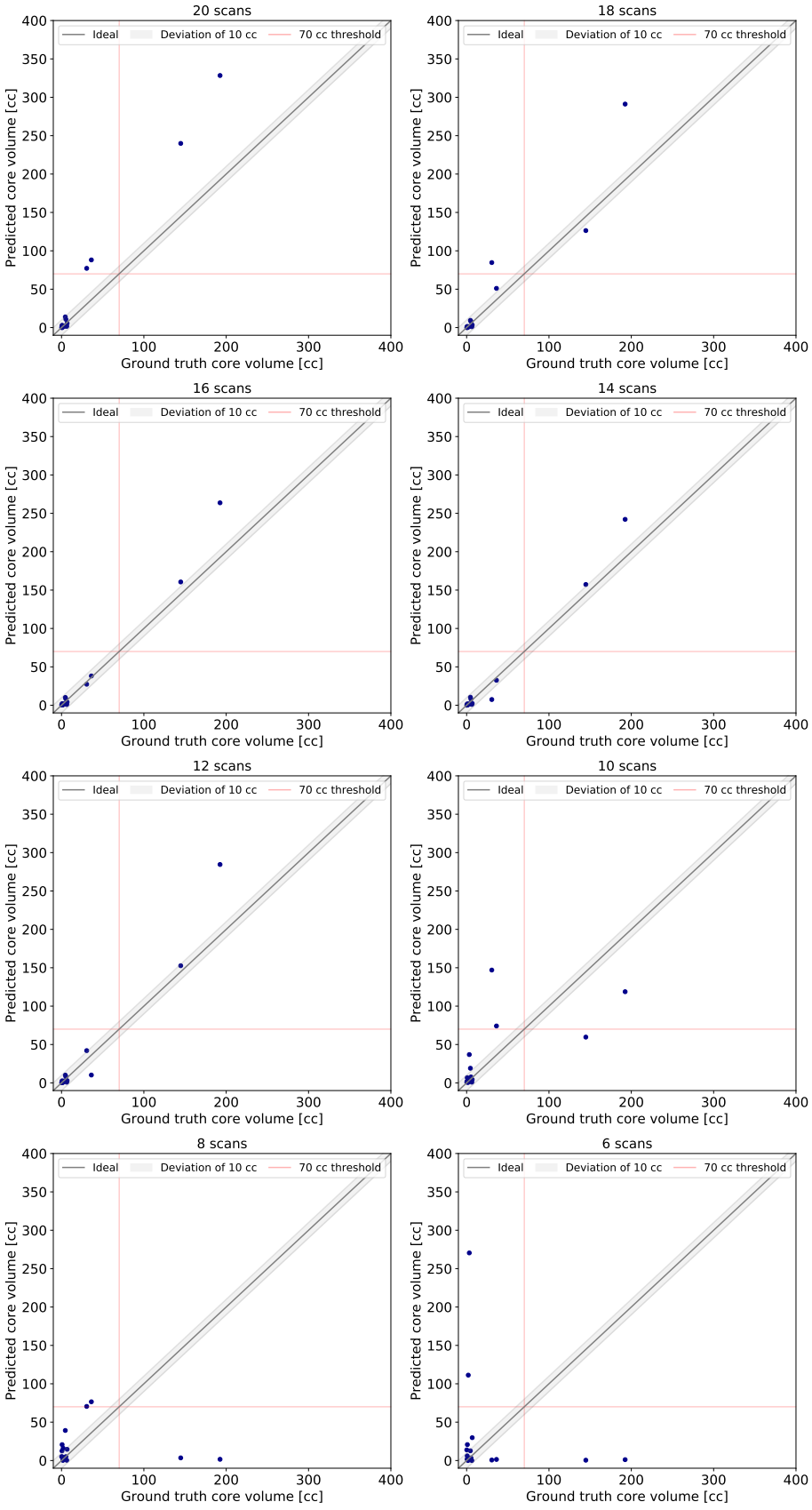



Figure 6.26: Overview of the impact of removing scans on the prediction of the core volume by a random forest using intensities as features.

For the situation with 10 scans remaining, the approach of removing every other one is considered as well. This plot, together with the one from the case where 10 scans are removed as explained above, are added in Appendix C.3. The result is slightly better, with two more cases where the treatment is predicted correctly, getting to a total of 15 out of 16. Regarding the number of patients with the deviation of core volume transcending 10 cc, there are now only four instead of six. Combined with the accuracies, this method proves to be better.
The localization error and relative difference in core volume are considered in Figure 6.27. The inaccuracy on the position gradually starts increasing with fewer scans. The deviation in volume becomes large from 10 scans onwards.

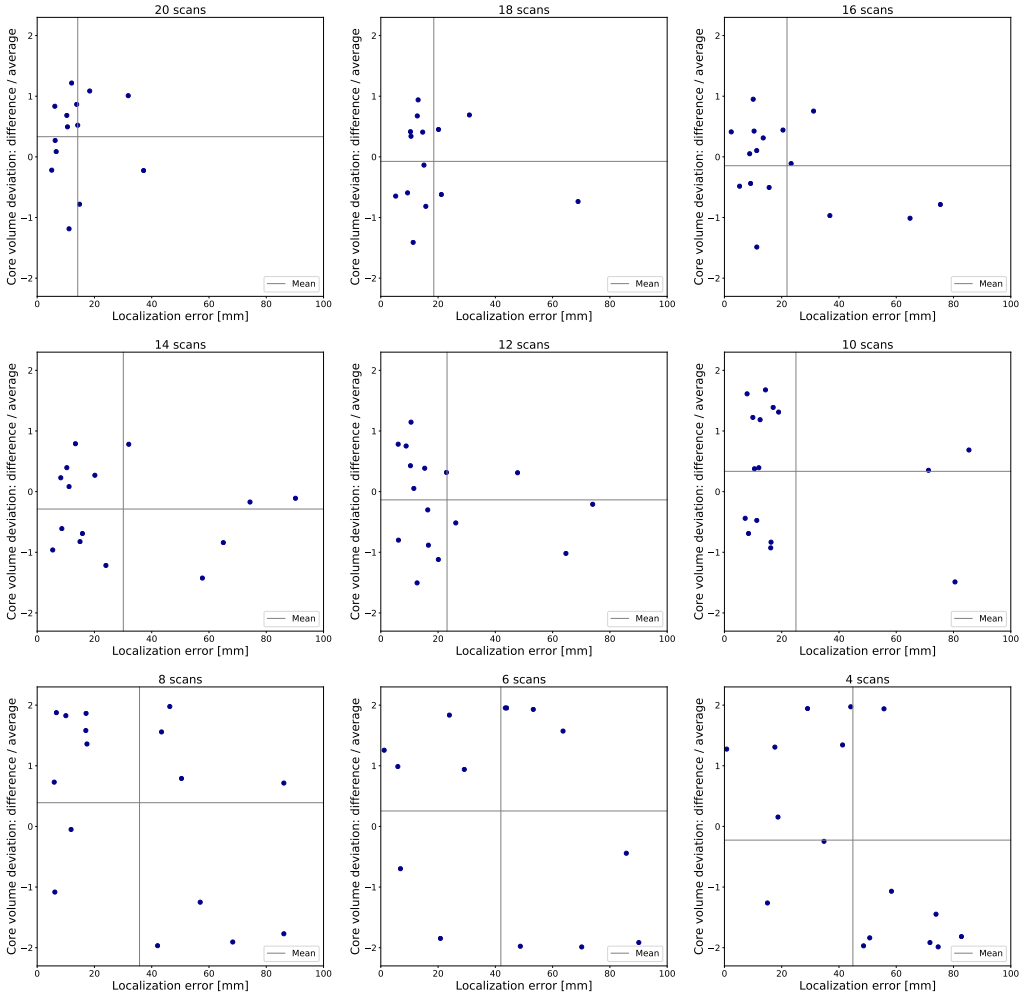



Figure 6.27: Overview of the impact of removing scans on the localization error and relative difference in core volume by a random forest using intensities as features.

### 6.3.2 Intensity derived features

Also for these binary core segmentations, the effect of adding intensity derived features is assessed, following the workflow depicted in Figure 6.28. It is possible that the extra information about the neighborhood of each voxel ensures a better performance with less scans. This hypothesis is investigated in the following paragraphs, using the same evaluation metrics as before.

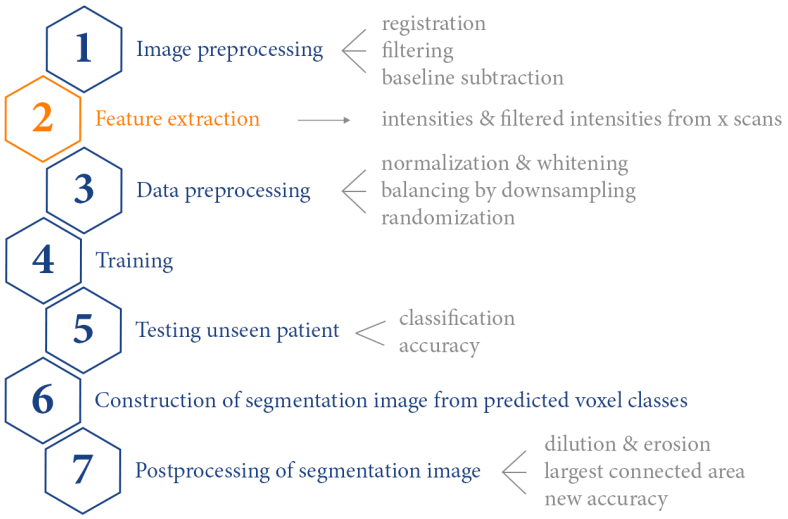



Figure 6.28: Workflow for prediction of core using the intensities and filtered intensities of $x$ scans as features with $x$ the number of scans considered.

The impact of removing scans is summarized in Figure 6.29. The prediction accuracy gradually decreases with the number of scans. Opposed to the results of the random forest using only the intensities as features, the accuracy of the postprocessed predictions no longer stays above 90 %. Boxplots of the accuracies per step are provided in Figure 6.30 while the ones for the postprocessed cases are depicted in Figure 6.31. For both, the distributions are similar to the initial situation up until 14 scans. With a further reduction, the deviations become larger. Also, the postprocessed cases are generally a bit less dispersed than the ones before postprocessing.
Using ten scans still provides an average accuracy of 87.01 % and 91.09 % after postprocessing. When every other scan is removed these are: 90.47 % and 95.42 % respectively. The latter method is thus slightly better.

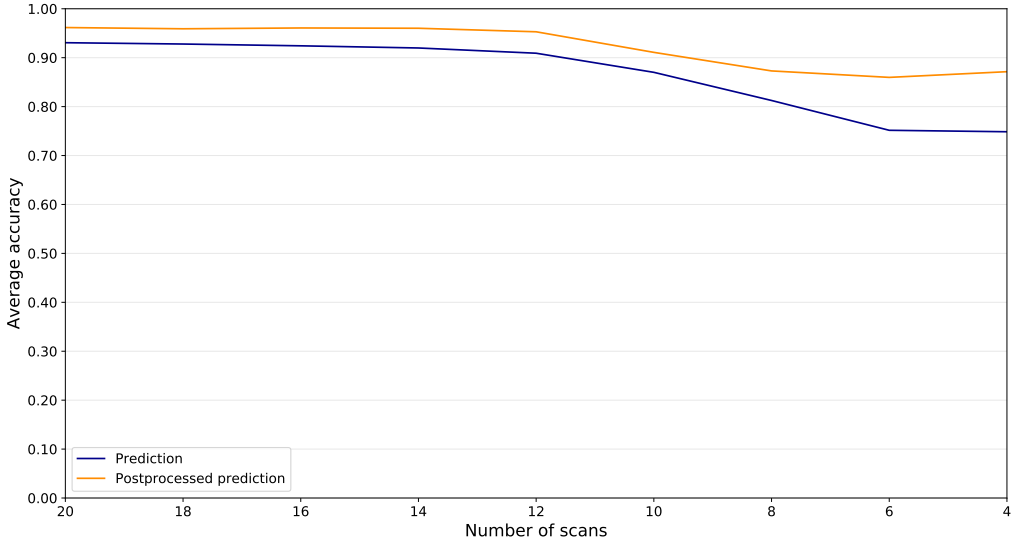



Figure 6.29: Influence of removing scans on the accuracy of a random forest using intensities as features, together with intensity derived features.

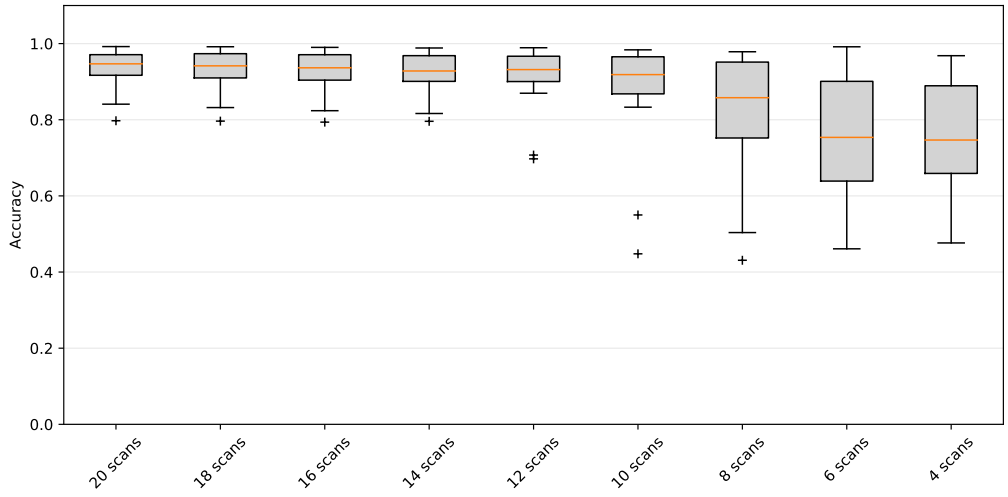



Figure 6.30: Boxplots showing the distribution of the accuracies for the leave-one-out experiments for the segmentation of the core per step of reducing scans. The random forest uses additional intensity derived features.

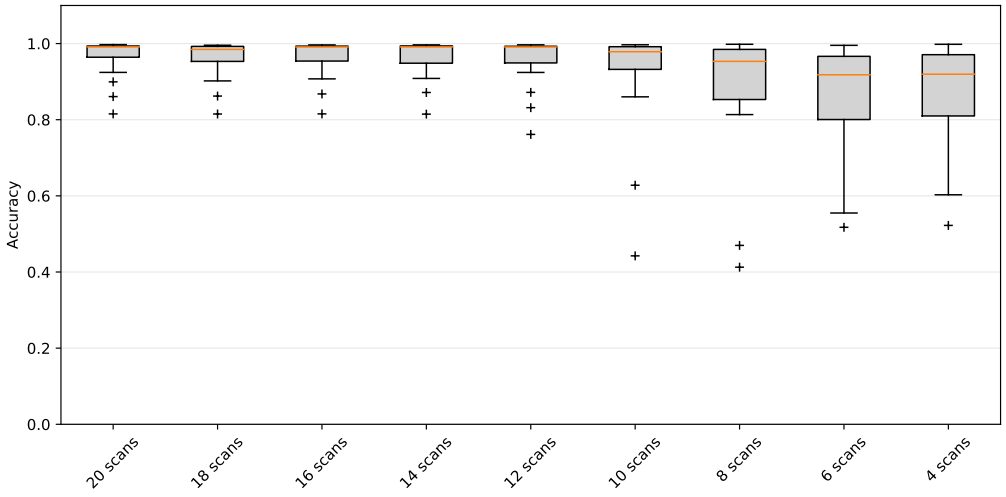



Figure 6.31: Boxplots showing the distribution of the postprocessed accuracies for the leave-one-out experiments for the segmentation of the core per step of reducing scans. The random forest uses additional intensity derived features.

Again, the impact of reducing the number of scans on the prediction of the core volume is assessed as well. The results are summarized in Figure 6.32. The performance stays comparable up until 12 scans. For ten scans, still there are only two cases where the treatment is not predicted correctly, but there are more patients for whom the deviation of core volume is bigger than 10 cc. From there onwards, the performance gets worse with respect to both parameters.

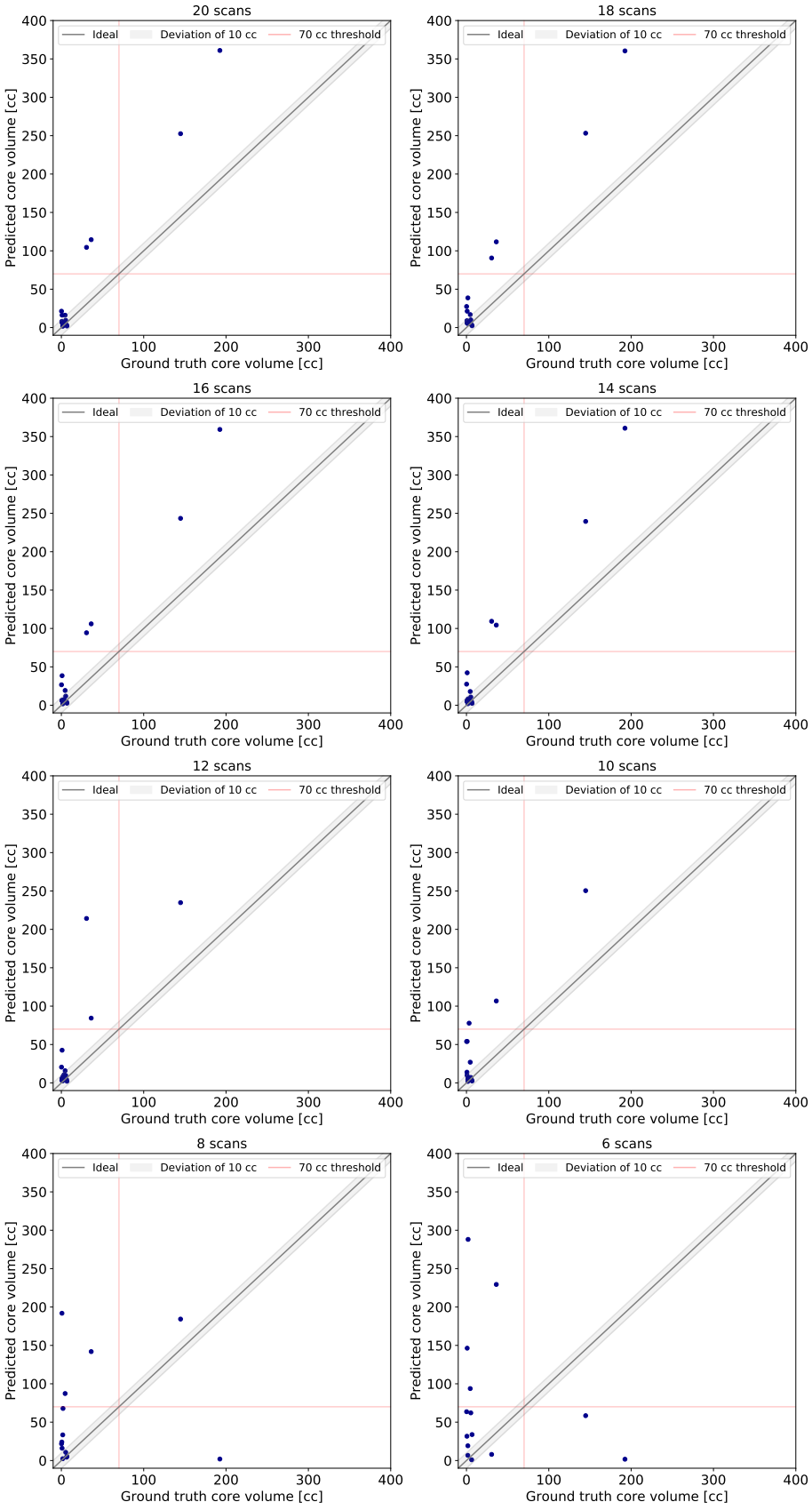



Figure 6.32: Overview of the impact of removing scans on the prediction of the core volume by a random forest using intensities as features, together with intensity derived features.

For 10 scans remaining, also the approach of removing every other one is considered. These plots are added in Appendix C.3. There is now one more patients with a correct prediction of the treatment, adding up to 14 out of 16. Also, there are two more patients without an excessive deviation of the predicted core volume from the ground truth. So, also for these features, this method of reducing scans proves to be better.
The localization error and relative difference in core volume are considered in Figure 6.33. From 10 scans onwards both errors become bigger for the majority of the patients.

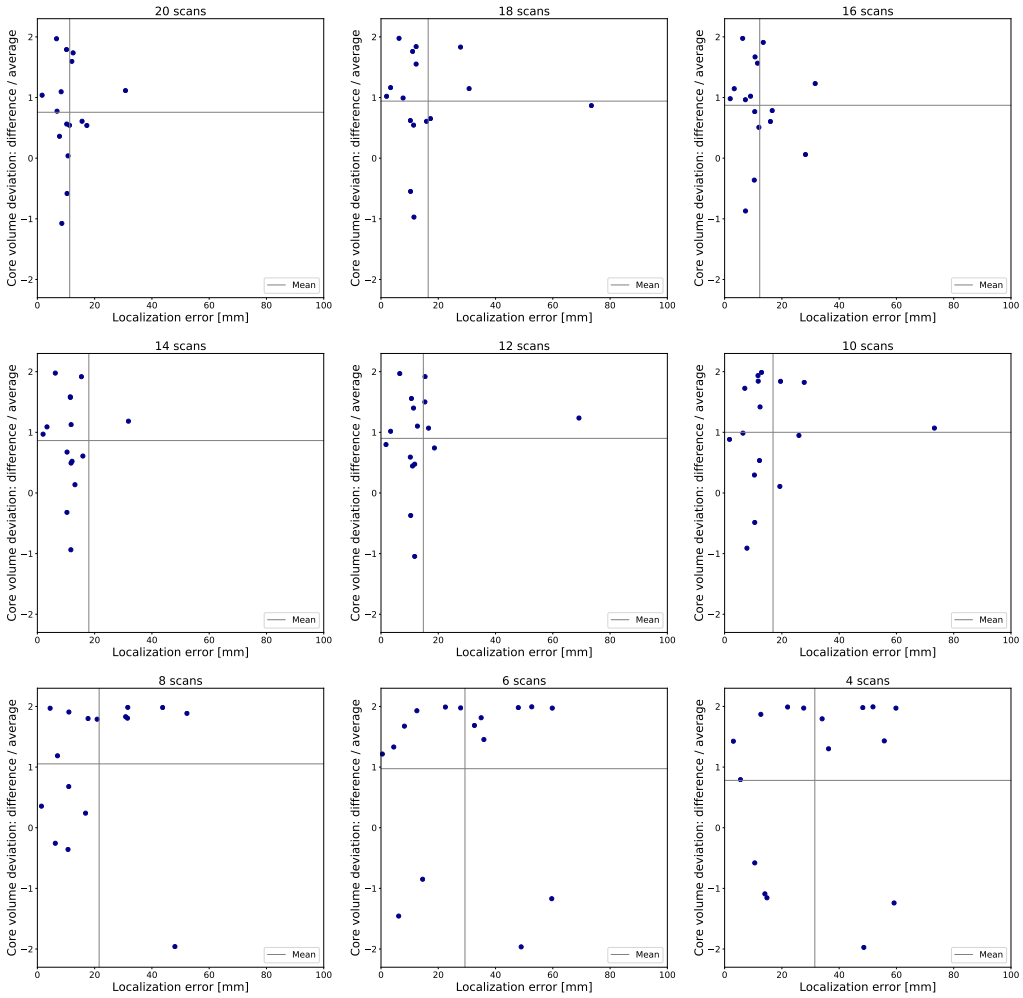



Figure 6.33: Overview of the impact of removing scans on the localization error and relative difference in core volume by a random forest using additional intensity derived features.

## 6.4 Conclusion

In this chapter the binary classifications were examined. The results for penumbra and healthy were slightly better than with the three-class problem. However, the fact that there were still patients with very bad segmentations suggest that the dataset is too small to obtain a generalized model. It is also possible a more complex classifier, like an ANN, is needed. For the core segmentations, the results were significantly better than before with respect to all considered metrics.
In line with the conclusions of the previous chapter, adding the intensity derived features allows the random forest to achieve higher accuracies. Postprocessing does not increase these values as much because the model already finds more connected regions. This is not the case when only the intensities are used as features and here the postprocessing allows the accuracy to rise about 13 % by removing the scattered misclassifications. When also taking into account the prediction of the core volume, the performance is better when only using the intensities. This is due to the extension of some segmentations into extra slices. The model obtaining more interconnected regions, has more voxels continuing in these extra slices and these are not reduced that much by the morphological operations. Besides this, the classifications are generally very good for both methods. The preference may be given to only using the intensities as features because similar results can be achieved with a lower computational demand.
The results when reducing the amount of scans, indicate that twenty scans are not needed to acquire all the necessary information. Ten to twelve scans should be sufficient, especially when the dataset can be expanded to include more examples of each class. Some of the outliers that are still present, are expected to perform better with a more generalized model. For keeping only ten time points, removing every other scan was observed to be the best option.

# Chapter 7

# Conclusion

### Goal

This thesis focused on finding a method to optimize the treatment of patients suspected of having a stroke. Currently, a perfusion CT is the golden standard to visualize the cerebral blood flow and derive perfusion maps from which treatment decisions can be made. This is, however, a complex procedure requiring a recent scanner and trained personnel. Therefore, the patient cannot always be transported to the nearest hospital. Since in the treatment of stroke time is essential, this is not an optimal situation. Furthermore, in the perfusion CT protocol generally 20 scans are taking. This imposes high requirements on the scanner, leads to a relatively high radiation dose for the patient and might not be the best way the use the little time there is for the treatment. In this dissertation it was investigated whether or not machine learning can be used to acquire knowledge from perfusion CT images regarding the characteristics of a perfusion deficit and regarding the aspects separating core from penumbra. Moreover, it was analyzed if there is a possibility to reduce the number of scans without losing necessary information. At the time this was written, no studies were found in literature researching this.

### Feasibility

Since it was the first time this dataset was used for the purpose of machine learning, a feasibility study was performed in Chapter 4. First, a simplified case was studied, selecting optimal voxels from different patients. Next, all the voxels in the brain were included, leaving one slice out for testing. This was done on intrapatient basis in order not to have problems with generalizing the model. Both small experiments confirmed that there is potential in using machine learning algorithm on these data.

### Results

In Chapter 5, a more extensive study was performed to analyze the ability of random forests to predict the location and size of core and penumbra. Some reliable predictions were obtained, endorsing the hypothesis. However, there were also bad results where the random forest failed to make an adequate classification. Probably the dataset of 44 patients was not large and diverse enough to obtain a truly general model. From these,

sixteen patients presented both classes and another eight contained only a penumbra region. The cerebral perfusion is very complex, differs between people and both core and penumbra can appear at any location and can have any shape or size. To include enough examples of every class, a larger dataset is needed. Also, the linear decision boundaries constructed by the random forest might not be sufficient to tackle this problem. A more complex classifier, like an artificial neural network, will be more suited for this.
In Chapter 6 the focus shifted towards a binary classification of the core. The lesion size is an important factor in the treatment decision. The random forest clearly performed better here. Both big and small lesions were found with a good estimate of their size for the majority of the patients.
For both the three-class and binary problem, the reduction of scans in time was analyzed. From this can be concluded that 20 scans are not needed to collect the necessary information. Depending on the features used, a reduction keeping only 10 to 12 scans is possible without major errors in the segmentations. The best option to retain ten scans is to remove every other time point.
From all the observations, it can be concluded that the pipeline from Figure 7.1 is the optimal one for predictions regarding the three-class problem, as well as predicting the core or best treatment option. Additional intensity derived features resulted in better initial performances. However, after postprocessing the accuracies when only intensities were used as features were comparable, while the estimation of the lesion size was closer to the ground truth. Together with the fact that adding extra features causes an increase in the computational demand, it can be concluded that the intensities should be employed, combined with a postprocessing step. In this research, reducing to ten scans generated some outliers. However, it is assumed that these would disappear once the dataset is general enough.

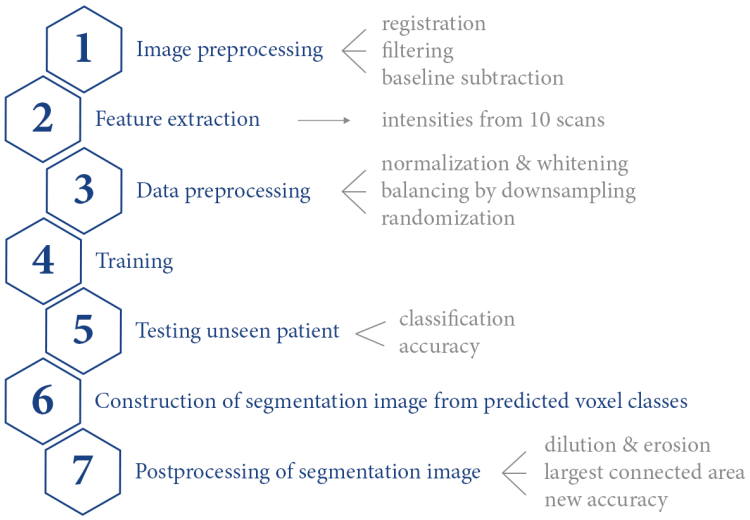



Figure 7.1: Optimal workflow.

### Limitations

A first limitation in this study was the definition of the ground truth labels. There is no consensus in literature about the optimal parameters and thresholds to derive core and penumbra from perfusion maps. Therefore, the labels were extracted using settings that resulted in reasonable segmentations for all patients. However, this might not correspond completely to the situation in reality.
Another limitation was the fact that this was the first time these data were provided as input for machine learning algorithms. Hence, it was chosen to focus on the necessary preprocessing, the feasibility and on machine learning techniques other than deep learning. This with the aim to enhance our understanding of the data and the classification problem.

### Future research

For future research on this topic, the construction of the ground truth labels should be improved. Including the follow-up diffusion-weighted MRI for the determination of the core, would already be more accurate. In this thesis, however, this was not possible because not for every patient such an MRI was available.
Next, the dataset should be expanded to obtain a more generalized model. It is expected that the outliers will perform better as well as the prediction of penumbra and healthy tissue. If indeed penumbra results become comparable to those for core segmentation, this can also be taken into account for the prediction of the therapeutic decision. It would be interesting to develop a model that can assess all the requirements on which the decision is based and provide the doctor with a suggestion for the method that gives the highest likelihood of a good clinical outcome.
Furthermore, the use of complexer classifiers, such artificial neural networks, should be investigated. When the correct parameters and settings are acquired, these are expected to achieve results superior to the random forest. These are indeed no longer constrained by the linear decision boundaries and can tackle more complicated problems.

### Final conclusion

In summary, this thesis has proven that machine learning has the potential to replace the complicated derivation of perfusion maps by gaining knowledge from perfusion CT images. It can provide the doctors with the needed information on the location and size of core and penumbra in stroke patients. Also, the number of scans in the perfusion CT protocol could be reduced, imposing a lower demand on the scanner, lowering the radiation dose for the patient as well as the time before the actual re-perfusion. The great advantage regarding this method is that it could be used in any hospital, meaning the patient could be transported to the nearest one and not the nearest one with a stroke unit. More research is needed, including a bigger and more diverse dataset as well as more complex classifiers. It is assumed these would only improve the results achieved here. If, in the end, machine learning models can replace the current treatment option, stroke patients will get a better chance of a good clinical outcome and a higher quality of life afterwards.

# Bibliography


[1] Safe: Stroke Alliance For Europe. URL http://www.safestroke.eu/about-stroke/.

[2] OnHealth. URL https://www.onhealth.com/content/1/stroke_signs_causes.

[3] Jeffrey L. Saver. Time is brain - Quantified. *Stroke*, 37(1):263–266, 2006. ISSN 00392499. doi: 10.1161/01.STR.0000196957.55928.ab.

[4] Stroke Shield Foundation. URL https://strokeshieldfoundation.org/our-research-focus/ischemic-versus-hemorrhagic-stroke-types/.

[5] David S. Liebeskind. Collateral circulation, 2003. ISSN 00392499.

[6] J.E. Delgado Almandoz, P. W. Schaefer, N. P. Forero, J. R. Falla, R. G. Gonzalez, and J. M. Romero. Diagnostic accuracy and yield of multidetector CT angiography in the evaluation of spontaneous intraparenchymal cerebral hemorrhage. *American Journal of Neuroradiology*, 30(6):1213–1221, 2009. ISSN 01956108. doi: 10.3174/ajnr.A1546.

[7] Z. Zaheer, T. Robinson, and A.K. Mistri. Thrombolysis in acute ischaemic stroke: an update. *Therapeutic Advances in Chronic Disease*, 2(2):119–131, 2011. ISSN 2040-6223. doi: 10.1177/2040622310394032.

[8] S.A. El Tawil and K.W. Muir B. Cme Cerebrovascular Disease. *Clinical Medicine*, 17(2):161–5, 2017. ISSN 10989064. doi: 10.1055/s-0036-1585078.

[9] Healthline. URL https://www.healthline.com/health/cerebral-angiography.

[10] Ctisus. URL http://www.ctisus.com/teachingfiles/neuro/320294.

[11] W.L. Bi, P. Brown, M. Abolfotoh, O. Al-mefty, S.M. Jr, and I.F. Dunn. Utility of dynamic computed tomography angiography in the preoperative evaluation of skull base tumors. 123(July):1–8, 2015. doi: 10.3171/2014.10.JNS141055.

[12] F. Scalzo and D.S. Liebeskind. Perfusion Angiography in Acute Ischemic Stroke. *Computational and Mathematical Methods in Medicine*, 2016, 2016. ISSN 17486718. doi: 10.1155/2016/2478324.

[13] S. P. Sourbron and D. L. Buckley. Tracer kinetic modelling in MRI: estimating perfusion and capillary permeability. *Physics in Medicine and Biology*, 57(2):R1–R33, 2012. ISSN 0031-9155. doi: 10.1088/0031-9155/57/2/R1.

[14] B. Romain, L. Rouet, D. Ohayon, O. Lucidarme, F. D'alché-Buc, and V. Letort. Parameter estimation of perfusion models in dynamic contrast-enhanced imaging: a unified framework for model comparison. *Medical Image Analysis*, 35:360–374, 2016. doi: 10.1016/j.media.2016.07.008.

[15] A. Fieselmann, M. Kowarschik, A. Ganguly, J. Hornegger, and R. Fahrig. Deconvolution-based CT and MR brain perfusion measurement: Theoretical model revisited and practical implementation details, 2011. ISSN 16874188.

[16] M. Wintermark, A.E. Flanders, B. Velthuis, R. Meuli, M. Van Leeuwen, D. Goldsher, C. Pineda, J. Serena, I. Van Der Schaaf, A. Waaijer, J. Anderson, G. Nesbit, I. Gabriely, V. Medina, A. Quiles, S. Pohlman, M. Quist, P. Schnyder, J. Bogousslavsky, W.P. Dillon, and S. Pedraza. Perfusion-CT assessment of infarct core and penumbra: Receiver operating characteristic curve analysis in 130 patients suspected of acute hemispheric stroke. *Stroke*, 37(4):979–985, 2006. ISSN 00392499. doi: 10.1161/01.STR.0000209238.61459.39.

[17] A. Bivard, C. Levi, N. Spratt, and M. Parsons. Perfusion CT in Acute Stroke: A Comprehensive Analysis of Infarct and Penumbra. *Radiology*, 2013. ISSN 0033-8419. doi: 10.1148/radiol.12120971.

[18] Y. Yu, Q. Han, Xi. Ding, Q. Chen, K. Ye, S. Zhang, S. Yan, B.C.V. Campbell, M.W. Parsons, S. Wang, and M. Lou. Defining Core and Penumbra in Ischemic Stroke: A Voxel- and Volume-Based Analysis of Whole Brain CT Perfusion. *Scientific Reports*, 2016. ISSN 20452322. doi: 10.1038/srep20932.

[19] X. Huang, D. Kalladka, B. Cheripelli, F. Moreton, and K.W. Muir. The Impact of CT Perfusion Threshold on Predicted Viable and Nonviable Tissue Volumes in Acute Ischemic Stroke. *Journal of Neuroimaging*, 2017. ISSN 15526569. doi: 10.1111/jon.12442.

[20] C.S. Kidwell, M. Wintermark, D.A. De Silva, T.J. Schaewe, R. Jahan, S. Starkman, T. Jovin, J. Hom, M. Jumaa, J. Schreier, J. Gornbein, D.S. Liebeskind, J.R. Alger, and J.L. Saver. Kidwell2013. *Stroke*, 44(1):73–79, 2012. doi: 10.1161/STROKEAHA.112.670034.

[21] G. Mair and J.M. Wardlaw. Imaging of acute stroke prior to treatment: current practice and evolving techniques. *Br J Radiol*, 87, 2014. doi: 10.1259/bjr.20140216.

[22] B.K. Menon, C.D. D'esterre, E.M. Qazi, M. Almekhlafi, L. Hahn, A.M. Demchuk, and M. Goyal. Multiphase cT angiography: A New Tool for the Imaging Triage of Patients with Acute Ischemic Stroke 1. *Radiology*, 275(2):510–520, 2015. ISSN 0033-8419. doi: 10.1148/radiol.15142256.

[23] S. E. Beyer, K. M. Thierfelder, L. Von Baumgarten, M. Rottenkolber, F. G. Meinel, H. Janssen, B. Ertl-Wagner, M. F. Reiser, and Wieland H. Sommer. Strategies of collateral blood flow assessment in ischemic stroke: Prediction of the follow-up infarct volume in conventional and dynamic CTA. *American Journal of Neuroradiology*, 36 (3):488–494, 2015. ISSN 1936959X. doi: 10.3174/ajnr.A4131.

[24] D. Byrne, G. Sugrue, E. Stanley, J .P. Walsh, S. Murphy, E .C. Kavanagh, and P .J. Macmahon. Improved Detection of Anterior Circulation Occlusions: The 'Delayed Vessel Sign' on Multiphase CT Angiography. *American Journal of Neuroradiology*, 38(10):1911–1916, 2017. doi: 10.3174/ajnr.A5317.

[25] A. I. Calleja, E. Cortijo, P. García-Bermejo, R. D. Gómez, S. Pérez-Fernández, J. M. del Monte, M. F. Muñoz, R. Fernández-Herranz, and J. F. Arenillas. Collateral circulation on perfusion-computed tomography-source images predicts the response to stroke intravenous thrombolysis. *European Journal of Neurology*, 20(5):795–802, 2013. ISSN 13515101. doi: 10.1111/ene.12063.

[26] M. Ernst, N. D. Forkert, L. Brehmer, G. Thomalla, S. Siemonsen, J. Fiehler, and A. Kemmling. Prediction of infarction and reperfusion in stroke by flowand volume-weighted collateral signal in MR angiography. *American Journal of Neuroradiology*, 36(2):275–282, 2015. ISSN 1936959X. doi: 10.3174/ajnr.A4145.

[27] V. Nambiar, S. I. Sohn, M. A. Almekhlafi, H. W. Chang, S. Mishra, E. Qazi, M. Eesa, A. M. Demchuk, M. Goyal, M. D. Hill, and Bijoy K. Menon. CTA collateral status and response to recanalization in patients with acute ischemic stroke. *American Journal of Neuroradiology*, 39(5):884–890, 2014. ISSN 1936959X. doi: 10.3174/ajnr.A3817.

[28] E. Martinon, P.H. Lefevre, P. Thouant, G.V. Osseby, F. Ricolfi, and A. Chavent. Collateral circulation in acute stroke: Assessing methods and impact: A literature review. *Journal of Neuroradiology*, 41(2):97–107, 2014. ISSN 17730406. doi: 10.1016/j.neurad.2014.02.001.

[29] B.K. Menon, B. O'Brien, A. Bivard, N.J. Spratt, A.M. Demchuk, F. Miteff, X. Lu, C. Levi, and M.W. Parsons. Assessment of Leptomeningeal Collaterals Using Dynamic CT Angiography in Patients with Acute Ischemic Stroke. *Journal of Cerebral Blood Flow & Metabolism*, 33(3):365–371, 2013. ISSN 0271-678X. doi: 10.1038/jcbfm.2012.171.

[30] A.M.J. Frölich, S.L. Wolff, M.N. Psychogios, E. Klotz, R. Schramm, K. Wasser, M. Knauth, and P. Schramm. Time-resolved assessment of collateral flow using 4D CT angiography in large-vessel occlusion stroke. *European Radiology*, 24(2):390–396, 2014. ISSN 09387994. doi: 10.1007/s00330-013-3024-6.

[31] B. K. Menon, E. E. Smith, J. Modi, S. K. Patel, R. Bhatia, T. W J Watson, M. D. Hill, A. M. Demchuk, and Mayank Goyal. Regional leptomeningeal score on CT angiography predicts clinical and imaging outcomes in patients with acute anterior circulation occlusions. *American Journal of Neuroradiology*, 32(9):1640–1645, 2011. ISSN 01956108. doi: 10.3174/ajnr.A2564.

[32] I. R. Van Den Wijngaard, G. Holswilder, M. J H Wermer, J. Boiten, A. Algra, D. W J Dippel, J. W. Dankbaar, B. K. Velthuis, A. M M Boers, C. B L M Majoie, and M. A A Van Walderveen. Assessment of collateral status by dynamic ct angiography in acute mca stroke: Timing of acquisition and relationship with final infarct volume. *American Journal of Neuroradiology*, 2016. ISSN 1936959X. doi: 10.3174/ajnr.A4746.

[33] Amy Y.X. Yu, C. Zerna, Z. Assis, J.K. Holodinsky, P.A. Randhawa, M. Najm, M. Goyal, B.K. Menon, A.M. Demchuk, S.B. Coutts, and M.D. Hill. Multiphase CT angiography increases detection of anterior circulation intracranial occlusion. *Neurology*, 87(6):609–616, 2016. ISSN 1526632X. doi: 10.1212/WNL.0000000000002951.

[34] C.D. D'Esterre, A. Trivedi, P. Pordeli, M. Boesen, S. Patil, S. Hwan Ahn, M. Najm, E. Fainardi, J.J.S. Shankar, M. Rubiera, M.A. Almekhlafi, J. Mandzia, A.V. Khaw, P. Barber, S. Coutts, M.D. Hill, A.M. Demchuk, T. Sajobi, N.D. Forkert, M. Goyal, T.Y. Lee, and B.K. Menon. Regional Comparison of Multiphase Computed Tomographic Angiography and Computed Tomographic Perfusion for Prediction of Tissue Fate in Ischemic Stroke. *Stroke*, 48(4):939–945, 2017. ISSN 15244628. doi: 10.1161/STROKEAHA.116.015969.

[35] O. Volny, P. Cimflova, P. Kadlecova, P. Vanek, J. Vanicek, B.K. Menon, and R. Mikulik. Single-Phase Versus Multiphase CT Angiography in Middle Cerebral Artery Clot Detection—Benefits for Less Experienced Radiologists and Neurologists. *Journal of Stroke and Cerebrovascular Diseases*, 26(1):19–24, 2017. ISSN 15328511. doi: 10.1016/j.jstrokecerebrovasdis.2016.08.023.

[36] A. Flores, M. Rubiera, M. Ribó, J. Pagola, D. Rodriguez-Luna, M. Muchada, S. Boned, L. Seró, E. Sanjuan, P. Meler, D. Carcámo, E. Santamarina, A. Tomassello, M. Lemus, P. Coscojuela, and C.A. Molina. Poor Collateral Circulation Assessed by Multiphase Computed Tomographic Angiography Predicts Malignant Middle Cerebral Artery Evolution after Reperfusion Therapies. *Stroke*, 46(11):3149–3153, 2015. ISSN 15244628. doi: 10.1161/STROKEAHA.115.010608.

[37] A. García-Tornel, V. Carvalho, S. Boned, A. Flores, D. Rodríguez-Luna, J. Pagola, M. Muchada, E. Sanjuan, P. Coscojuela, J. Juega, N. Rodriguez-Villatoro, B. Menon, M. Goyal, M. Ribó, A. Tomasello, C.A. Molina, and Marta Rubiera. Improving the Evaluation of Collateral Circulation by Multiphase Computed Tomography Angiography in Acute Stroke Patients Treated with Endovascular Reperfusion Therapies. *Interventional Neurology*, 5(3-4):209–217, 2016. ISSN 1664-9737. doi: 10.1159/000448525.

[38] E. Tong, J. Patrie, S. Tong, A. Evans, P. Michel, A. Eskandari, and M. Wintermark. Time-resolved CT assessment of collaterals as imaging biomarkers to predict clinical outcomes in acute ischemic stroke. *Neuroradiology*, 59(11):1101–1109, 2017. doi: 10.1007/s00234-017-1914-z.

[39] P. Bentley, J. Ganesalingam, A. Lalani Carlton Jones, K. Mahady, S. Epton, P. Rinne, P. Sharma, O. Halse, A. Mehta, and D. Rueckert. Prediction of stroke thrombolysis outcome using CT brain machine learning. *NeuroImage: Clinical*, 4: 635–640, 2014. doi: 10.1016/j.nicl.2014.02.003.

[40] D. Zikic, B. Glocker, E. Konukoglu, A. Criminisi, C. Demiralp, J. Shotton, O. M. Thomas, T. Das, R. Jena, and S. J. Price. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 15(Pt 3):369–76, 2012. ISSN 16113349. doi: 10.1007/978-3-642-33454-2{_}46.

[41] J. Mitra, P. Bourgeat, J. Fripp, S. Ghose, S. Rose, O. Salvado, A. Connelly, B. Campbell, S. Palmer, G. Sharma, S. Christensen, and L. Carey. Lesion segmentation from multimodal MRI using random forest following ischemic stroke. *NeuroImage*, 98: 324–335, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.04.056.

[42] E. Lee, Y. Kim, N. Kim, and D. Kang. Deep into the Brain: Artificial Intelligence in Stroke Imaging. *Journal of Stroke*, 2017. ISSN 2287-6391. doi: 10.5853/jos.2017.02054.

[43] O. Maier, C. Schröder, N.D. Forkert, T. Martinetz, and H. Handels. Classifiers for ischemic stroke lesion segmentation: A comparison study. *PLoS ONE*, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0145118.

[44] D. Pustina, H. Branch Coslett, P.E. Turkeltaub, N. Tustison, M.F. Schwartz, and B. Avants. Automated segmentation of chronic stroke lesions using LINDA: Lesion Identification with Neighborhood Data Analysis HHS Public Access. *Hum Brain Mapp*, 37(4):1405–1421, 2016. doi: 10.1002/hbm.23110.

[45] L. Chen, P. Bentley, and D. Rueckert. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *NeuroImage: Clinical*, 2017. ISSN 22131582. doi: 10.1016/j.nicl.2017.06.016.

[46] M. Scherer, J. Cordes, A. Younsi, Y.A. Sahin, M. Götz, M. Möhlenbruch, C. Stock, J. Bösel, A. Unterberg, K. Maier-Hein, and B. Orakcioglu. Development and Validation of an Automatic Segmentation Algorithm for Quantification of Intracerebral Hemorrhage. *Stroke*, 2016. ISSN 0039-2499. doi: 10.1161/STROKEAHA.116.013779.

[47] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990. ISSN 01628828. doi: 10.1109/34.56205.

[48] A. Fränzle, J. Hillengass, and R. Bendl. Spinal focal lesion detection in multiple myeloma using multimodal image features. 2015. doi: 10.1117/12.2081990.

[49] B.C.V. Campbell, P.J. Mitchell, B. Yan, M.W. Parsons, S. Christensen, L. Churilov, R.J. Dowling, H. Dewey, M. Brooks, F. Miteff, C. Levi, M. Krause, T.J. Harrington, K.C. Faulder, B.S. Steinfort, T. Kleinig, R. Scroop, S. Chryssidis, A. Barber, A. Hope, M. Moriarty, B. Mcguinness, A.A. Wong, A. Coulthard, T. Wijeratne, A. Lee, J. Jannes, J. Leyden, T.G. Phan, W. Chong, MiE. Holt, R.V. Chandra, C.F. Bladin, M. Badve, H. Rice, L. de Villiers, H. Ma, P.M. Desmond, G.A. Donnan, and S.M. Davis. A multicenter, randomized, controlled study to investigate extending the time for thrombolysis in emergency neurological deficits with intra-arterial therapy (EXTEND-IA). *International Journal of Stroke*, 9(1):126–132, 2014. ISSN 17474930. doi: 10.1111/ijs.12206.

[50] G.W. Albers, M.P. Marks, S. Kemp, S. Christensen, J.P. Tsai, S. Ortega-Gutierrez, R.A. McTaggart, M.T. Torbey, M. Kim-Tenser, T. Leslie-Mazwi, A. Sarraj, Scott E. Kasner, Sameer A. Ansari, Sharon D. Yeatts, S. Hamilton, M. Mlynash, Jeremy J. Heit, G. Zaharchuk, S. Kim, J. Carrozzella, Y.Y. Palesch, A.M. Demchuk, R. Bammer, P.W. Lavori, J.P. Broderick, and M.G. Lansberg. Thrombectomy for Stroke at 6 to 16 Hours with Selection by Perfusion Imaging. *New England Journal of Medicine*, 378(8):708–718, 2018. ISSN 0028-4793. doi: 10.1056/NEJMoa1713973.

# Appendix A

# Results feasibility study

## A.1 Manual selection of data

A full overview of the results from the feasibility experiments on manually selected points is given in the following tables. The header 'Accuracy cv' refers to the accuracy through 10-fold cross-validation and 'Accuracy test' is the accuracy on the test set that was not included in the training. Table A.1 shows the results for the SVM classifier with various values for both C and gamma.

| **Parameters** | | **Test patient** | **Accuracy cv* on training set** | **Accuracy on test set** |
|---|---|---|---|---|
| C = 1 | $\gamma$ = auto | 1 | 0.9992 | 0.8438 |
| | | 2 | 0.9992 | 0.9863 |
| | | 3 | 0.9996 | 0.8838 |
| | | 4 | 0.9992 | 0.6113 |
| | | 5 | 0.9996 | 0.8150 |
| | | 6 | 0.9990 | 0.8925 |
| | | 7 | 0.9994 | 0.7288 |
| **average** | | | **0.9993** | **0.8230** |
| C = 10 | $\gamma$ = auto | 1 | 1.0000 | 0.8325 |
| | | 2 | 0.9996 | 0.9913 |
| | | 3 | 0.9998 | 0.8875 |
| | | 4 | 0.9996 | 0.6038 |
| | | 5 | 1.0000 | 0.8238 |
| | | 6 | 1.0000 | 0.8913 |
| | | 7 | 1.0000 | 0.7413 |
| **average** | | | **0.9999** | **0.8245** |
| C = 100 | $\gamma$ = auto | 1 | 1.0000 | 0.8350 |
| | | 2 | 0.9998 | 0.9938 |
| | | 3 | 1.0000 | 0.8875 |
| | | 4 | 0.9998 | 0.6025 |
| | | 5 | 1.0000 | 0.8200 |
| | | 6 | 1.0000 | 0.8925 |

| | | | | |
|---|---|---|---|---|
| | | 7 | 1.0000 | 0.7413 |
| **average** | | | **0.9999** | **0.8246** |
| C = 1000 | $\gamma$ = auto | 1 | 1.0000 | 0.8350 |
| | | 2 | 0.9998 | 0.9938 |
| | | 3 | 1.0000 | 0.8875 |
| | | 4 | 0.9998 | 0.6025 |
| | | 5 | 1.0000 | 0.8200 |
| | | 6 | 1.0000 | 0.8925 |
| | | 7 | 1.0000 | 0.7413 |
| **average** | | | **0.9999** | **0.8246** |
| C = 100 | $\gamma$ = 0.01 | 1 | 0.9992 | 0.8213 |
| | | 2 | 0.9985 | 0.9900 |
| | | 3 | 0.9990 | 0.8813 |
| | | 4 | 0.9988 | 0.6688 |
| | | 5 | 0.9992 | 0.8225 |
| | | 6 | 0.9992 | 0.8650 |
| | | 7 | 0.9992 | 0.7575 |
| **average** | | | **0.9990** | **0.8295** |
| C = 100 | $\gamma$ = 0.1 | 1 | 1.0000 | 0.8713 |
| | | 2 | 1.0000 | 0.9850 |
| | | 3 | 1.0000 | 0.8850 |
| | | 4 | 0.9998 | 0.5525 |
| | | 5 | 1.0000 | 0.7938 |
| | | 6 | 1.0000 | 0.9188 |
| | | 7 | 1.0000 | 0.7575 |
| **average** | | | **1.0000** | **0.8234** |
| C = 100 | $\gamma$ = 1 | 1 | 0.9946 | 0.5000 |
| | | 2 | 0.9975 | 0.5138 |
| | | 3 | 0.9881 | 0.5000 |
| | | 4 | 0.9817 | 0.5000 |
| | | 5 | 0.9779 | 0.5000 |
| | | 6 | 0.9940 | 0.5000 |
| | | 7 | 0.9952 | 0.5013 |
| **average** | | | **0.9899** | **0.5021** |

Table A.1: Results from manually selected points in the brain for the SVM classifier with various values for C and $\gamma$ (*cv = cross-validation).

Table A.2 summarizes the results for the decision tree classifier with various values for both the maximum depth and minimum number of leaf samples.

| Parameters | | Test patient | Accuracy cv* on training set | Accuracy on test set |
|---|---|---|---|---|
| Max depth* = none | MLS = 1 | 1 | 0.9631 | 0.7113 |
| | | 2 | 0.9606 | 0.9575 |
| | | 3 | 0.9606 | 0.8875 |
| | | 4 | 0.9792 | 0.7375 |
| | | 5 | 0.9698 | 0.7513 |
| | | 6 | 0.9546 | 0.8150 |
| | | 7 | 0.9669 | 0.7000 |
| **average** | | | **0.9650** | **0.7943** |
| Max depth* = 100 | MLS = 1 | 1 | 0.9631 | 0.7113 |
| | | 2 | 0.9606 | 0.9575 |
| | | 3 | 0.9606 | 0.8875 |
| | | 4 | 0.9792 | 0.7375 |
| | | 5 | 0.9698 | 0.7513 |
| | | 6 | 0.9546 | 0.8150 |
| | | 7 | 0.9669 | 0.7000 |
| **average** | | | **0.9650** | **0.7943** |
| Max depth* = 10 | MLS = 1 | 1 | 0.9604 | 0.6188 |
| | | 2 | 0.9558 | 0.9588 |
| | | 3 | 0.9581 | 0.8963 |
| | | 4 | 0.9771 | 0.7313 |
| | | 5 | 0.9656 | 0.7525 |
| | | 6 | 0.9579 | 0.8525 |
| | | 7 | 0.9617 | 0.6763 |
| **average** | | | **0.9624** | **0.7838** |
| Max depth* = none | MLS = 10 | 1 | 0.9502 | 0.7000 |
| | | 2 | 0.9408 | 0.9688 |
| | | 3 | 0.9460 | 0.9250 |
| | | 4 | 0.9592 | 0.7325 |
| | | 5 | 0.9571 | 0.8313 |
| | | 6 | 0.9477 | 0.8850 |
| | | 7 | 0.9608 | 0.7000 |
| **average** | | | **0.9517** | **0.8204** |
| M depth* = none | MLS = 100 | 1 | 0.9040 | 0.5988 |
| | | 2 | 0.8952 | 0.9300 |
| | | 3 | 0.8923 | 0.9163 |
| | | 4 | 0.9315 | 0.7738 |
| | | 5 | 0.9223 | 0.7850 |
| | | 6 | 0.9071 | 0.8288 |
| | | 7 | 0.9223 | 0.7563 |
| **average** | | | **0.9107** | **0.7984** |

Table A.2: Results from manually selected points in the brain for the decision tree classifier with various values for the maximum depth and the minimum number of leaf samples (*cv = cross-validation, max depth = maximum depth, MLS = minimum leaf samples).

From the previous results it is clear that no constraint on the maximum depth and a minimum number of leaf samples equal to 10 gives the best results. Therefore these settings will be used in the random forests as well. A range of number of estimators is tested and the results are presented in Table A.3

| **Parameters** | **Test patient** | **Accuracy cv* on training set** | **Accuracy on test set** |
|---|---|---|---|
| N = 10 | 1 | 0.9713 | 0.7775 |
| | 2 | 0.9642 | 0.9750 |
| | 3 | 0.9683 | 0.9275 |
| | 4 | 0.9813 | 0.7013 |
| | 5 | 0.9679 | 0.7988 |
| | 6 | 0.9710 | 0.9038 |
| | 7 | 0.9742 | 0.7375 |
| **average** | | **0.9712** | **0.8316** |
| N = 20 | 1 | 0.9706 | 0.7900 |
| | 2 | 0.9733 | 0.9875 |
| | 3 | 0.9748 | 0.9350 |
| | 4 | 0.9838 | 0.7163 |
| | 5 | 0.9725 | 0.7963 |
| | 6 | 0.9752 | 0.8975 |
| | 7 | 0.9748 | 0.7413 |
| **average** | | **0.9750** | **0.8377** |
| N = 50 | 1 | 0.9733 | 0.8113 |
| | 2 | 0.9704 | 0.9938 |
| | 3 | 0.9765 | 0.9325 |
| | 4 | 0.9860 | 0.7163 |
| | 5 | 0.9769 | 0.7863 |
| | 6 | 0.9752 | 0.9000 |
| | 7 | 0.9773 | 0.7500 |
| **average** | | **0.9765** | **0.8414** |
| N = 100 | 1 | 0.9744 | 0.8363 |
| | 2 | 0.9715 | 0.9913 |
| | 3 | 0.9760 | 0.9275 |
| | 4 | 0.9854 | 0.7200 |
| | 5 | 0.9779 | 0.7875 |
| | 6 | 0.9748 | 0.8988 |
| | 7 | 0.9783 | 0.7450 |
| **average** | | **0.9769** | **0.8438** |
| N = 200 | 1 | 0.9744 | 0.8463 |
| | 2 | 0.9733 | 0.9888 |
| | 3 | 0.9767 | 0.9338 |
| | 4 | 0.9858 | 0.7288 |
| | 5 | 0.9779 | 0.7888 |
| | 6 | 0.9763 | 0.9025 |
| | 7 | 0.9792 | 0.7525 |
| **average** | | **0.9776** | **0.8488** |

| | | | |
|---|---|---|---|
| N* = 500 | 1 | 0.9760 | 0.8463 |
| | 2 | 0.9752 | 0.9888 |
| | 3 | 0.9767 | 0.9313 |
| | 4 | 0.9860 | 0.7313 |
| | 5 | 0.9800 | 0.7900 |
| | 6 | 0.9771 | 0.8988 |
| | 7 | 0.9796 | 0.7488 |
| **average** | | **0.9787** | **0.8479** |

Table A.3: Results from manually selected points in the brain for the random forest classifier with various values for the number of estimators (*cv = cross-validation, N = number of estimators).

## A.2 Incorporation of the whole brain

The prediction accuracies per slice by a random forest, trained on the remaining slices, are provided in Table A.4.

| **Slice** | **Accuracy** | **Accuracy postprocessed** |
|---|---|---|
| 1 | 0.6912 | 0.7532 |
| 2 | 0.7240 | 0.8101 |
| 3 | 0.7063 | 0.8219 |
| 4 | 0.6851 | 0.7814 |
| 5 | 0.7170 | 0.7788 |
| 6 | 0.7272 | 0.8105 |
| 7 | 0.7294 | 0.8025 |
| 8 | 0.7374 | 0.7996 |
| 9 | 0.7549 | 0.8226 |
| 10 | 0.7554 | 0.8269 |
| 11 | 0.7554 | 0.8124 |
| 12 | 0.7137 | 0.7830 |
| 13 | 0.7424 | 0.7929 |
| 14 | 0.7287 | 0.7434 |
| 15 | 0.6264 | 0.6366 |
| **average** | **0.7196** | **0.7851** |

Table A.4: Prediction accuracies per slice by a random forest, trained on the remaining slices.

# Appendix B

# Predicting core and penumbra

## B.1 Results of paired t-tests

To check the statistical significance of the difference between the four cases, a paired t-test is performed. The dataset for each experiment consists of the same patients. A significance level of 1 % is chosen, so $\alpha = 0.01$. Because more than two experiments are compared, the chance of a rare event increases. To compensate for this, the Bonferroni correction is applied, meaning $\alpha$ is divided by the number of hypotheses. So, the new significance threshold becomes $\alpha = 0.01/4 = 0.025$. The results are summarized in Table B.1.

| Pair | | p-value | Significant difference? |
|---|---|---|---|
| Intensities | PP* intensities | 0.03698 | no |
| Filtered intensities | PP* filtered intensities | 0.44134 | no |
| Intensities | Filtered intensities | $2.11681.10^{-17}$ | yes |
| PP* intensities | PP* filtered intensities | 0.02605 | no |

Table B.1: P-values of paired t-test to check the statistical difference between the accuracies of the three class segmentations (*PP = postprocessed).
With the Bonferroni correction: $\alpha = 0.025$.

The p-values for the different scan reduction steps are tabulated in Table B.2 and, for the postprocessed cases, Table B.3.

| **Pair** | | **p-value** |
|---|---|---|
| 20 scans | 18 scans | $3.83873.10^{-5}$ |
| 18 scans | 16 scans | 0.00998 |
| 16 scans | 14 scans | $8.17293.10^{-6}$ |
| 14 scans | 12 scans | $8.71833.10^{-6}$ |
| 12 scans | 10 scans | 0.00063 |
| 10 scans | 8 scans | 0.39981 |
| 8 scans | 6 scans | $3.92363.10^{-6}$ |
| 6 scans | 4 scans | 0.01205 |
| | | |
| 20 scans | 10 scans | $1.20172.10^{-8}$ |
| 20 scans | 8 scans | $3.64280.10^{-5}$ |
| 20 scans | 6 scans | $8.88770.10^{-8}$ |
| 20 scans | 4 scans | $1.68307.10^{-9}$ |

Table B.2: P-values of paired t-test to check the statistical difference between the accuracies of the 3-class segmentation for the different steps of reducing scans in case of intensities as features.

| **Pair** | | **p-value** |
|---|---|---|
| 20 scans | 18 scans | 0.00365 |
| 18 scans | 16 scans | 0.35554 |
| 16 scans | 14 scans | 0.54551 |
| 14 scans | 12 scans | 0.15193 |
| 12 scans | 10 scans | 0.60410 |
| 10 scans | 8 scans | 0.31127 |
| 8 scans | 6 scans | 0.00382 |
| 6 scans | 4 scans | 0.77957 |
| | | |
| 20 scans | 10 scans | 0.71849 |
| 20 scans | 8 scans | 0.44356 |
| 20 scans | 6 scans | 0.26415 |
| 20 scans | 4 scans | 0.42086 |

Table B.3: P-values of paired t-test to check the statistical difference between the accuracies of the 3-class segmentation for the different steps of reducing scans in case of intensities as features and postprocessing.

## B.2 Enlarged images

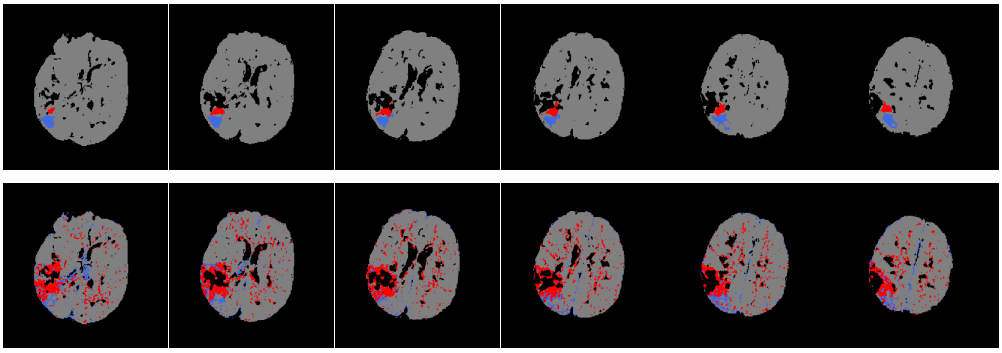


Figure B.1: Zoom on the most relevant slices of Figure 5.2

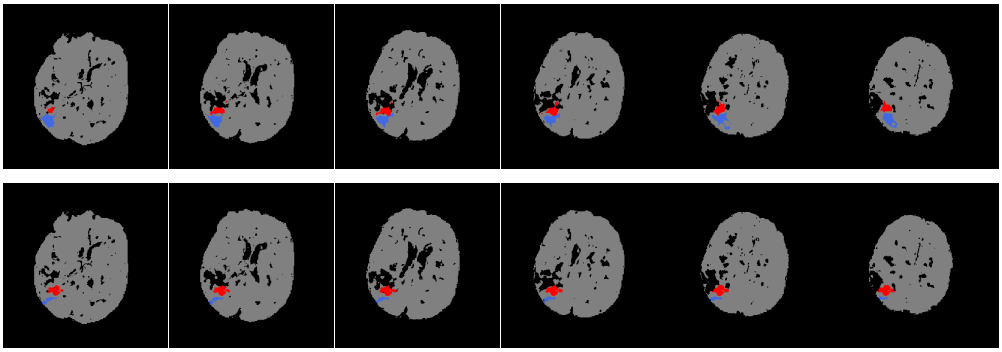


Figure B.2: Zoom on the most relevant slices of Figure 5.5

# Appendix C

# Predicting the therapeutic decision

## C.1 Results of paired t-tests

The statistical differences between the accuracies are checked with a pair t-test and these results are summarized in Table C.1.

| Pair | | p-value | Significant difference? |
|---|---|---|---|
| Intensities | PP* intensities | $1.41749.10^{-6}$ | yes |
| Filtered intensities | PP* filtered intensities | 0.12808 | no |
| Intensities | Filtered intensities | 0.00026 | yes |
| PP* intensities | PP* filtered intensities | 0.73464 | no |

Table C.1: P-values of paired t-test to check the statistical difference between the accuracies of the core segmentations (*PP = postprocessed).
With the Bonferroni correction: $\alpha = 0.025$.

The p-values for the different scan reduction steps are tabulated in Table C.2 and, for the postprocessed cases, Table C.3.

| **Pair** | | **p-value** |
|---|---|---|
| 20 scans | 18 scans | 0.00058 |
| 18 scans | 16 scans | 0.02278 |
| 16 scans | 14 scans | 0.06154 |
| 14 scans | 12 scans | 0.35544 |
| 12 scans | 10 scans | 0.015972 |
| 10 scans | 8 scans | 0.02742 |
| 8 scans | 6 scans | 0.02853 |
| 6 scans | 4 scans | 0.69837 |
| | | |
| 20 scans | 10 scans | 0.00356 |
| 20 scans | 8 scans | 0.00116 |
| 20 scans | 6 scans | 0.00013 |
| 20 scans | 4 scans | 1.08452 |

Table C.2: P-values of paired t-test to check the statistical difference between the accuracies of the core segmentation for the different steps of reducing scans in case of intensities as features.

| **Pair** | | **p-value** |
|---|---|---|
| 20 scans | 18 scans | 0.06782 |
| 18 scans | 16 scans | 0.06942 |
| 16 scans | 14 scans | 0.62732 |
| 14 scans | 12 scans | 0.64854 |
| 12 scans | 10 scans | 0.06172 |
| 10 scans | 8 scans | 0.25678 |
| 8 scans | 6 scans | 0.26452 |
| 6 scans | 4 scans | 0.92254 |
| | | |
| 20 scans | 10 scans | 0.17652 |
| 20 scans | 8 scans | 0.16905 |
| 20 scans | 6 scans | 0.21352 |
| 20 scans | 4 scans | 0.22918 |

Table C.3: P-values of paired t-test to check the statistical difference between the accuracies of the core segmentation for the different steps of reducing scans in case of intensities as features and postprocessing.

The p-values for the different scan reduction steps, in case of additional intensity derived features, are tabulated in Table C.4 and, for the postprocessed cases, Table C.5.

| **Pair** | | **p-value** |
|---|---|---|
| 20 scans | 18 scans | 0.03085 |
| 18 scans | 16 scans | 0.00675 |
| 16 scans | 14 scans | 0.00889 |
| 14 scans | 12 scans | 0.26935 |
| 12 scans | 10 scans | 0.05374 |
| 10 scans | 8 scans | 0.09870 |
| 8 scans | 6 scans | 0.11633 |
| 6 scans | 4 scans | 0.95759 |
| | | |
| 20 scans | 10 scans | 0.04163 |
| 20 scans | 8 scans | 0.00715 |
| 20 scans | 6 scans | 0.00051 |
| 20 scans | 4 scans | 0.00022 |

Table C.4: P-values of paired t-test to check the statistical difference between the accuracies of the core segmentation for the different steps of reducing scans in case of additional intensity derived features.

| **Pair** | | **p-value** |
|---|---|---|
| 20 scans | 18 scans | 0.33889 |
| 18 scans | 16 scans | 0.54669 |
| 16 scans | 14 scans | 0.48632 |
| 14 scans | 12 scans | 0.28540 |
| 12 scans | 10 scans | 0.10830 |
| 10 scans | 8 scans | 0.28269 |
| 8 scans | 6 scans | 0.76353 |
| 6 scans | 4 scans | 0.82309 |
| | | |
| 20 scans | 10 scans | 0.12317 |
| 20 scans | 8 scans | 0.05908 |
| 20 scans | 6 scans | 0.01982 |
| 20 scans | 4 scans | 0.02877 |

Table C.5: P-values of paired t-test to check the statistical difference between the accuracies of the core segmentation for the different steps of reducing scans in case of additional intensity derived features and postprocessing.

## C.2 Enlarged images

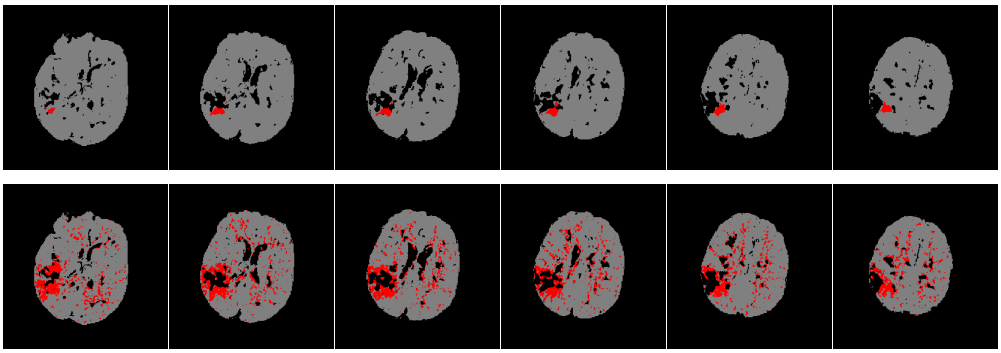


Figure C.1: Zoom on the most relevant slices of Figure 6.2

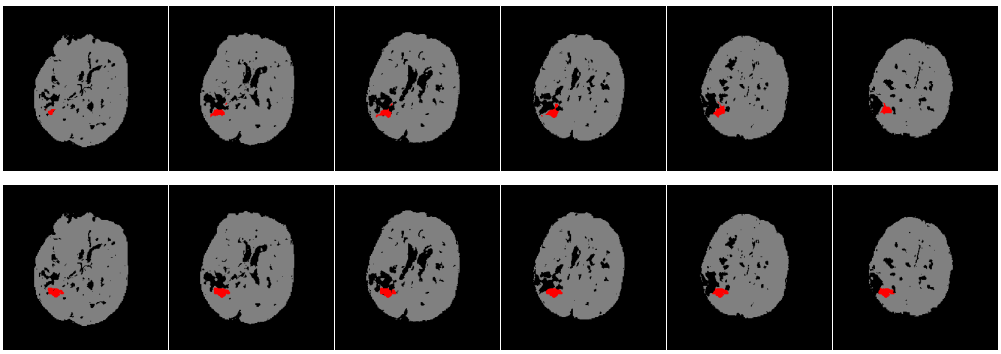


Figure C.2: Zoom on the most relevant slices of Figure 6.5

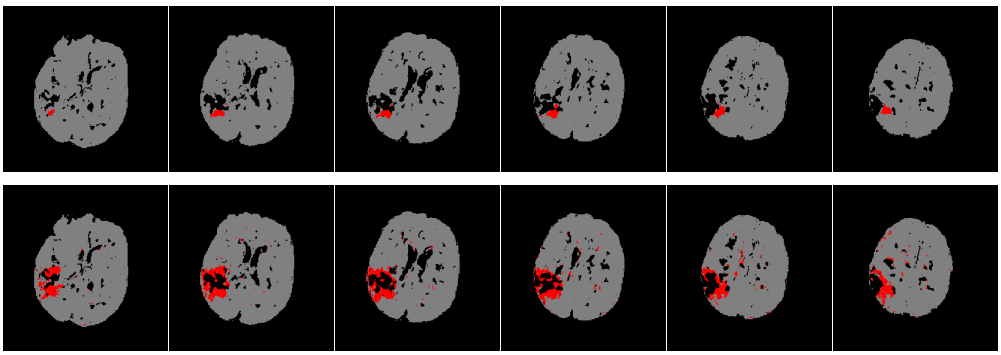


Figure C.3: Zoom on the most relevant slices of Figure 6.17

## C.3 Volume estimation for 10 scans

A comparison is made between removing 10 scans as explained in chapter 6 or by removing every other scan, for the intensities used as features in Figure C.4. The same is done in case of adding the intensity derived features in Figure C.5.

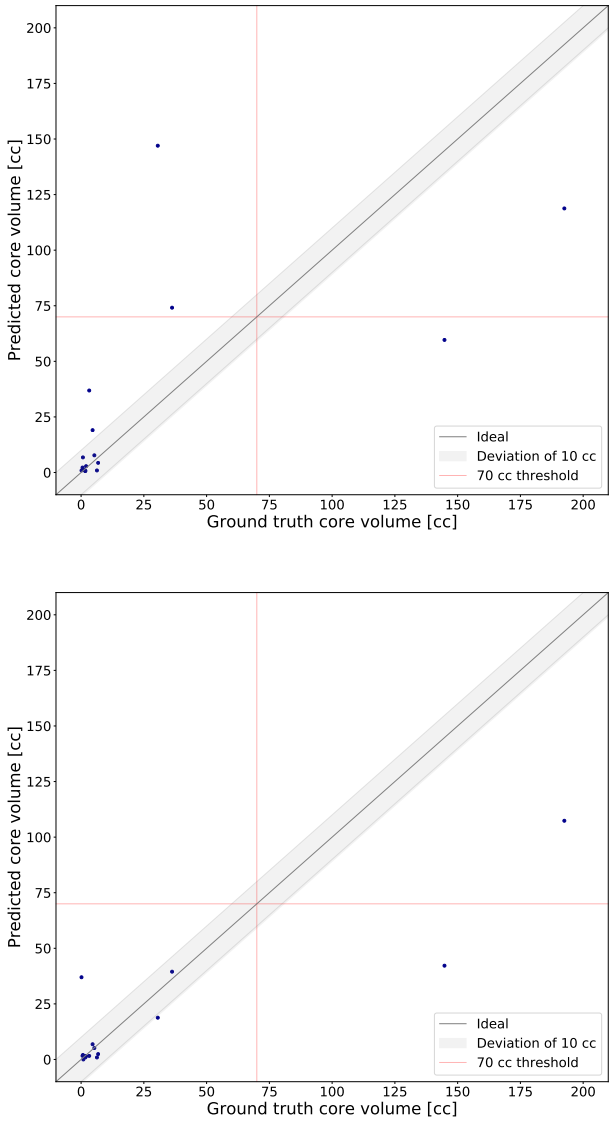



Figure C.4: Comparison of the influence of removing 10 scans as explained in chapter 6 (top) or by removing every other scan (bottom) on the prediction of the core volume by a random forest using intensities as features.

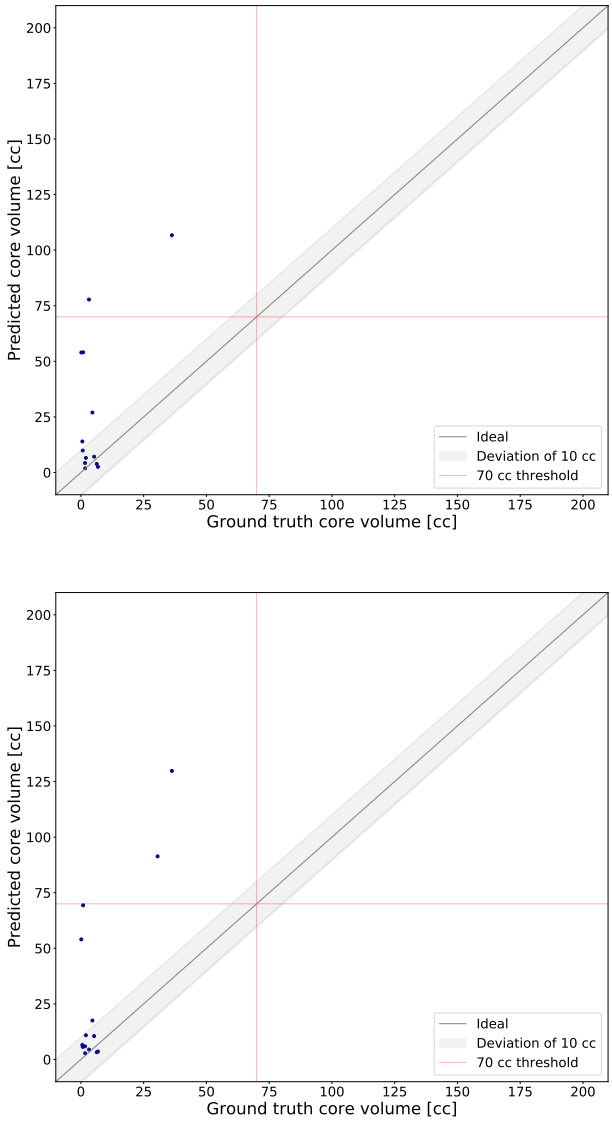



Figure C.5: Comparison of the influence of removing 10 scans as explained in chapter 6 (top) or by removing every other scan (bottom) on the prediction of the core volume by a random forest using intensities as features, together with intensity derived features.