Analysis of liver lesions in dynamic contrast enhanced MR images

Mariëlle J.A. Jansen

Analysis of liver lesions in dynamic contrast enhanced MR images

Mariëlle J.A. Jansen

PhD Thesis, Utrecht University, the Netherlands

Cover design: Luke Noothout
Thesis layout: Mariëlle Jansen
Printed by: Ridderprint | www.ridderprint.nl
ISBN: 987-90-393-7195-4

This research is supported by the project BENEFIT (Better Effectiveness aNd Efficiency by measuring and modelling of Interventional Therapy) and IMPACT (Intelligence based iMprovement of Personalized treatment And Clinical workflow supporT) both in the framework of the EU research programme ITEA (Information Technology for European Advancement).

Financial support from ChipSoft and Sectra Benelux is gratefully acknowledged.

Analysis of liver lesions in dynamic contrast enhanced MR images

Analyse van leverlaesies in dynamische contrastversterkte MR beelden

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
dinsdag 10 december 2019 des middags te 2.30 uur

door

Mariëlle Joanna Adriana Jansen

geboren op 22 oktober 1991
te Oss

Promotoren: Prof. dr. J.P.W. Pluim

Prof. dr. ir. M.A. Viergever

Copromotor: Dr. H.J. Kuijf

Voor mijn peetoom

Contents

1

General introduction

The liver is the largest internal organ in the body and performs many vital functions, including production of proteins and blood clotting factors, glycogen synthesis, bile production, and elimination of toxins.[1] Any dysfunctionality can severely affect the liver function and may thus have a large impact on patients' health and quality of life. Globally 840,000 people were diagnosed with primary liver cancer (hepatocellular carcinoma (HCC)) in 2018. Liver cancer is the fifth most common cancer worldwide and one of the major causes of cancer death, with 780,000 deaths in 2018.[2] Next to the HCC, malignant liver metastases prevail in the liver. Liver metastases are tumors that metastasize to the liver from other parts of the body, mainly from the gastrointestinal tract. Alongside these malignant lesions, a variety of benign liver lesions may occur in the liver, such as adenomas, hemangiomas, focal nodular hyperplasias (FNH), and cysts. These benign lesions prevail with a frequency up to 25% of patients.[3]

It is clinically very important that liver lesions are detected and differentiated as benign or malignant in time for proper treatment. There is a variety of possible treatments for malignant lesions, such as surgical resection, radiofrequency ablation, transarterial chemoembolization, chemotherapy, and in severe cases liver transplantation.[4] Treatment selection depends on the type of lesion, its location and the number of lesions.[5] To obtain this information, medical images are acquired through ultrasound, computed tomography (CT), or magnetic resonance imaging (MRI).

Although CT has long been the method of choice for detecting and monitoring liver lesions, MRI is becoming increasingly popular thanks to a better contrast between liver and lesion without the use of ionizing radiation.[6–8] In this thesis the focus will be on image analysis of liver MR images.

Magnetic resonance imaging

Different types of MR images can be obtained during a liver MR examination. Most common are: T1-weighted, T2-weighted, diffusion weighted, and dynamic contrast enhanced (DCE) images.

This thesis focuses on the use of DCE-MR images. DCE-MRI contains not only anatomical, but also functional information about tissue perfusion through the use of a contrast agent. A contrast agent increases the intensity of surrounding molecules on MR images. First a pre-contrast image is made after which the contrast agent is administered intravenously into the bloodstream. After contrast administration, several 3D images are acquired to follow the contrast agent and thereby also the

blood flow through the abdomen.

The liver has an inflow of blood through the hepatic artery as well as through the portal vein. This creates a complex system, resulting in different phases of contrast enhancement. The phases depend on the time between the administration of the contrast agent and the acquisition of the phase image. Figure 1 shows the six phase images of contrast enhancement in MR images; pre-contrast, early arterial phase, late arterial phase, hepatic/portal-venous phase, late portal venous/equilibrium phase, and late equilibrium phase. The contrast enhancement of different vessels and tissues in the six phases gives insight into the perfusion of the liver and its tissue properties.[9,10]

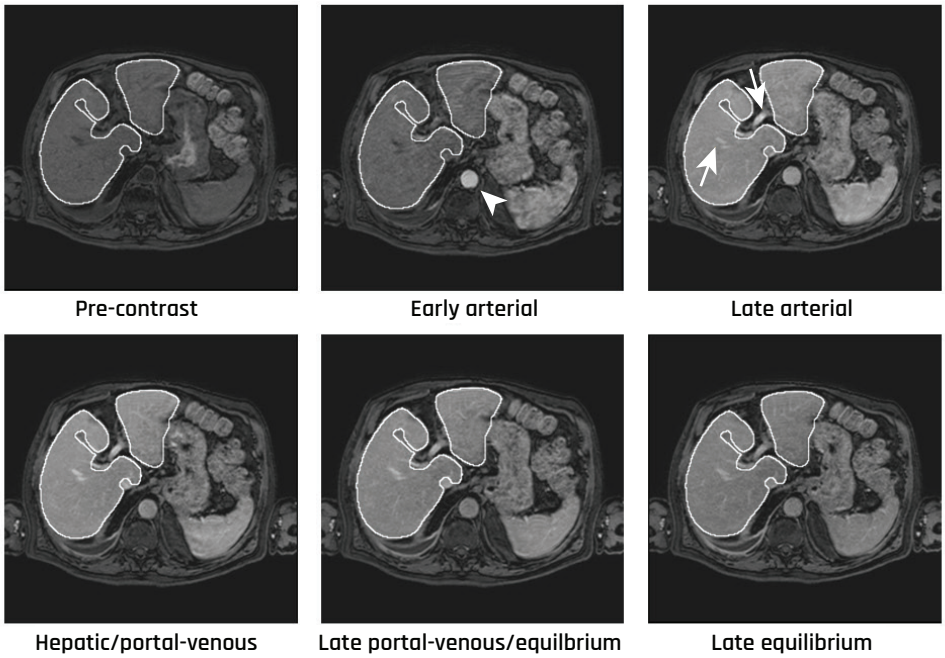


Figure 1 The six phases of dynamic contrast enhanced MRI. The liver is contoured in white. In the pre-contrast image, the MR intensities are shown without the influence of the contrast agent. In the early arterial phase the aorta contains a large amount of contrast agent (arrow head). In the late arterial phase the contrast agent arrives via the vessels (arrows) in the liver, brightening the liver parachyma. In the later phases the contrast agent washes out again.

The radiologist uses DCE-MRI, the other MR images, and clinical patient information to make a diagnosis. All sequences contain specific information that is useful for detecting and characterizing lesions. A high intensity on diffusion weighted MRI, for example, is a good indication of the presence of a lesion.[11] Enhancement of a lesion at a certain phase of the DCE-MRI series can be characteristic of malignancy.

Hyperintensity on a T2-weighted MR image is an indication of a cyst or hemangioma. The texture of a lesion on an MR image can also help with tissue characterization, such as rim enhancement in malignant lesions or heterogeneous enhancement of hemangiomas.[3,6,7,12] Figure 2 shows the arterial phase of DCE-MRI and the T2-weighted MR image of five different lesions. Image processing and image analysis can assist the radiologist in analyzing the images quickly and arriving at a diagnosis and a treatment plan.

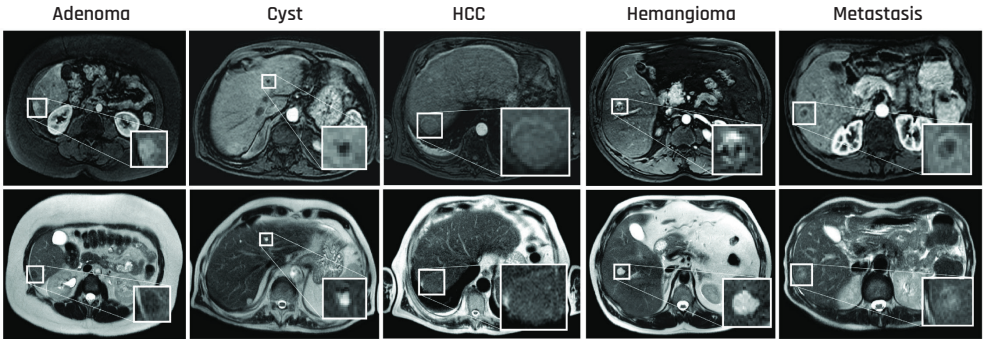


Figure 2 Examples of focal liver lesions. From left to right livers with an adenoma, cyst, HCC, hemangioma, and a metastasis from colorectal carcinoma origin are shown. The top row shows the arterial phases of DCE-MRI and the bottom row T2-weighted images. A zoom-in of the lesions is inserted.

Image registration

The liver is a somewhat flexible organ and is able to move a bit within the abdomen. This is important because the adjacent organs are in motion; a beating heart, intestinal peristalsis, and expanding lungs during inhalation. As a result, the liver does not always occupy the same position within the abdomen. These movements, mainly caused by breathing, can move the liver up or down by a few centimetres.[13] This motion makes it more difficult to find the same location in the liver in a different contrast phase or different MR image. Especially in DCE-MR images, since the acquisition can take up to ten minutes. Patients are therefore asked to hold their breath several times during the MR acquisition to limit the largest displacement factor. However, this does not lead to images without motion and motion artefacts. It is not always easy for patients to hold their breath at the same depth or for longer periods of time. The motion between images can be partially corrected after the MR acquisition by means of image registration.

Image registration is the process of spatial alignment of two or more images in such a manner that the same anatomical structures have the same coordinates in

all images.[14] This makes it easier to combine anatomical information from multiple images for image analysis.

During image registration, one of the images, *the moving image*, is deformed to align with the other image, *the fixed image*. The quality of alignment is measured using a cost function. Frequently used cost functions in medical image registration are mutual information[15] and normalized correlation coefficient, both similarity metrics based on the intensities of the two images. The cost function is iteratively optimized with respect to the deformation of the moving image. The deformation can consist of either linear transformations such as translation, rotation, scaling, and shearing, or nonlinear transformations, depending on the degrees of freedom given to the deformation. Figure 3 shows a schematic representation of the image registration framework.

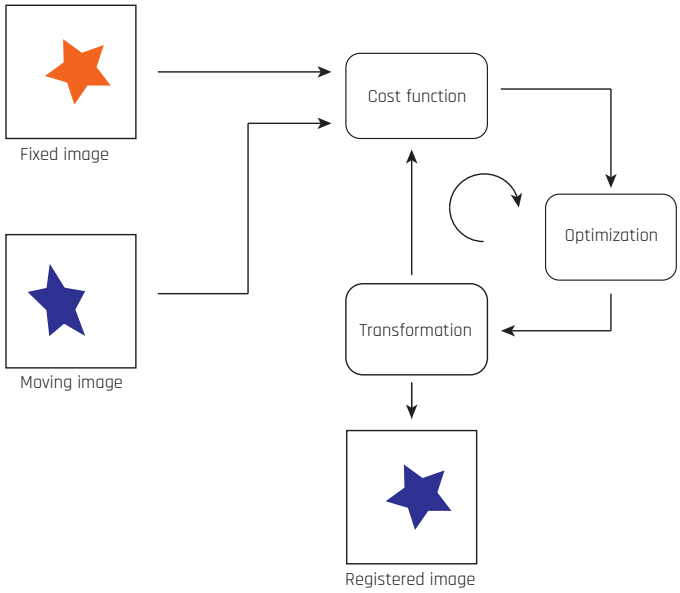


Figure 3 Schematic representation of the image registration framework. The degree of alignment is measured with the cost function. During the optimization step of the cost function is optimized by updating the transformation. This new transformation is applied to the moving image and again the degree of alignment is measured with the cost function and so on. This process is iterated until the specified number of iterations or the optimum in the cost function is reached

Especially nonlinear transformations are crucial for an accurate image registration between two abdominal images, since the liver is moving nonlinearly itself. Alignment between the different phases of DCE-MRI can be achieved as well as alignment between different MR sequences. This simplifies the examination of suspicious regions on different MR phases and sequences. Radiologists examine lesions on

different sequences by looking at two or more sequences next to each other. When the images are not registered, the radiologist has to find the slices that correspond to the same lesion. When the images are registered radiologists just looks at one slice and location, which is the same slice in all MR phases and sequences. In addition, the success of image analysis of multiple correlating images is depending on the alignment of these images. In Chapter 2 of this thesis, different image registration methods for DCE-MR images are compared and the method with the most satisfying results is used in the rest of this thesis.

Machine learning

Radiologists inspect the different MR sequences, localize lesions, and determine the type of lesion based on specific features. Machine learning algorithms can aid radiologists in detecting and classifying a lesion.

Machine learning is based on algorithms that can learn from data without being explicitly programmed to perform a task. For example a data set exists of two classes. In class 1 the objects are apples and in class 2 the objects are oranges. Features that describe these objects are for example color and roughness of the skin. A machine learning algorithm creates a split between the two classes based on the values of these features. If the object is red and has a smooth skin, the object is most likely an apple. On the other hand, if the object is orange and has a rough skin, the object is most likely an orange. The algorithm learns which feature values correspond with each class.

Machine learning can either be unsupervised or supervised. In unsupervised learning the data set is unlabeled i.e. no class is assigned to the object, and the algorithm has to find a pattern in the features to divide the data set in two or more groups. This is also called clustering. In supervised learning the data set is labelled and the algorithm learns a decision model to assign a class label to a new object based on previously seen objects with the corresponding features and classes. In this thesis only supervised machine learning algorithms are used.

Supervised machine learning algorithms are presented with a data set containing the features and the corresponding class labels, called the training set. The algorithm, called classifier, maps the features and learns which feature values are associated with each class. New data points are classified according to the learned decision model. Figure 4A shows an example of two easy separable classes with a linear classifier, i.e. the boundary between the two classes is linear. However in some cases, as shown in

Figure 4B, a linear classifier is not sufficient and not all data points in the data set are assigned the right class by the linear classifier. Nonlinear classifiers are needed to properly split the data in the two classes, see Figure 4C.

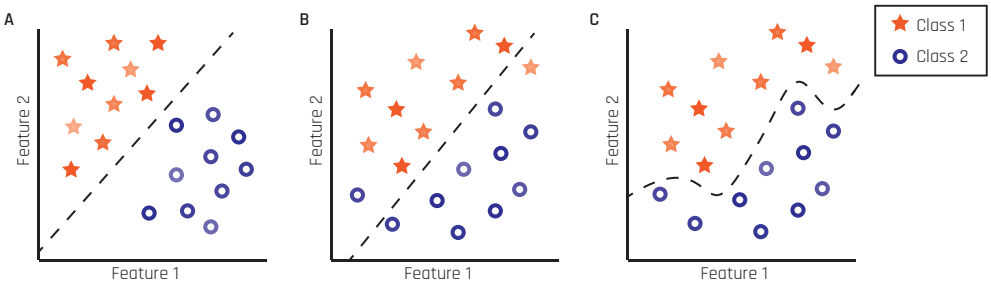


Figure 4 Scatterplots displaying the relation between feature 1 and feature 2 of two classes. A) A linear classifier that can separate the given data set. B) A linear classifier that cannot separate the given data set. C) Nonlinear classifier that can separate the given data set of B.

Classifiers, especially nonlinear classifiers have a risk of overfitting to the data. A classifier is overfitting, when it does not generalize well enough and is unable to assign the right class to new data points. This can occur if there is not enough data in the training set, if the data points do not represent the general population, or if a classifier has too many degrees of freedom.[16] In all cases the classifier fails to learn the correct boundary. Figure 5 shows an example of overfitting.

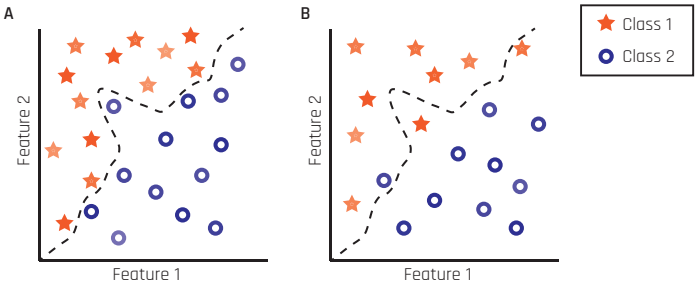


Figure 5 A) A nonlinear fitting of the classifier on the training set. B) The same decision boundary is applied to new data points. As a result of overfitting, the classifier is unable to assign the correct class in some cases.

In these examples only 2 features are shown, but many features can be extracted from an image, e.g. the intensity, roundness, or size of the object. However, not all features contain useful information for a classifier. For example if a classifier is designed to distinguish apples and oranges, the color of the object is a useful feature, since most oranges are orange and apples are red, yellow, or green, but not orange. The diameter of the object is a less useful feature, since apples and oranges can have a

similar diameter. If features with no useful information are presented to the classifier, the classifier might not be able to learn the right decision boundary. In addition, having too many features will increase the risk of overfitting. In order to present the classifier with a small number of useful features, feature selection is often done. A feature selection algorithm selects a number of features based on the information density of the features.

The feature selection procedure is mainly applied when the features are designed or handcrafted based on human interpretations of the classification task. In medical images, these features are often based on the characteristics radiologists are looking for, i.e. contrast uptake rate or mean intensity level in the lesion. In this thesis we used this approach to classify different focal liver lesions.

Another option is to learn the features from the input image directly instead of designing the features first. This can be done by using a deep learning algorithm. A deep learning algorithm does not only learn the features, but can also function as a classifier. The algorithm is trained by iteratively presenting the algorithm with an input image and the corresponding class label and through a number of layers with weighted connections the relation between the input and the output is learned. During each iteration the weights are updated to minimize the error of the output. This error is measured using a loss function, a function to describe the cost of the output for a given label. In medical image analysis, convolutional neural networks (CNN) are the most commonly used deep learning algorithms.[17] Figure 6 gives a schematic overview of the framework of conventional machine learning and deep learning.

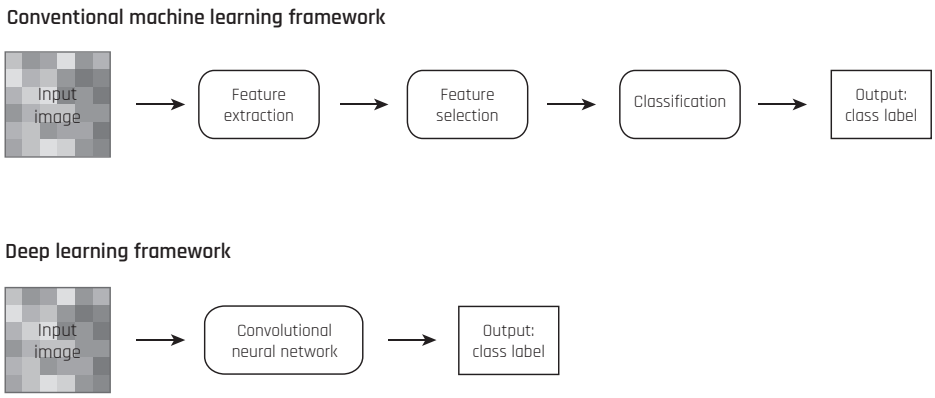


Figure 6 Overview of the framework for conventional machine learning (top row) and deep learning (bottom row).

Deep learning methods are able to learn relatively complex tasks and automatically extract the features most suited for the given tasks. The key of deep learning algorithms to solve these complex tasks is the large number of parameters. During the training of a CNN up to tens of millions of weights or parameters are optimized. As mentioned before, having many parameters increases the risk of overfitting. To reduce this risk, researchers have developed many techniques such as dropout[18], batch normalization[19], and data augmentation.

Despite these techniques, still a lot of data is needed. The necessary amount of good quality data is the biggest downside of deep learning methods in medical image analysis. The amount of necessary data is lower when the main task is split into multiple simpler tasks. Instead of learning an algorithm to provide the radiologist with a diagnosis based on an MR image, the task is split into a detection task, a segmentation task, and a lesion classification task. Both conventional machine learning algorithms and CNNs can be used for these different tasks. All three tasks are actually classification tasks, either on image level or on pixel level. Image level classification is for example classification of a lesion, by processing the part of image representing the lesion. Classification done on pixel level is mostly used for detection and segmentation tasks. In this thesis, deep learning algorithms are applied to segment the liver and detect lesions.

Thesis outline

The aim of this thesis is to explore and develop techniques to aid the radiologist in processing and analyzing MR images in order to obtain a proper diagnosis on hepatic focal lesions.

In **Chapter 2**, two different strategies for the registration of DCE-MR images are evaluated; pairwise and groupwise registration. These two methods were not evaluated before on clinical data with a low temporal resolution. In this chapter they are evaluated and compared based on quantitative metrics as well as on visual assessment by clinical experts.

In **Chapter 3**, the performance of three different input configurations of DCE-MR images for CNNs is studied for a liver segmentation task. The input configuration can have a large influence on the performance of a CNN. DCE-MR images have several phases that can be used as input images, but it is yet unknown which configuration performs best.

In **Chapter 4**, a two-step method based on CNNs to detect liver metastases on MR images is described. First, the liver is segmented using the optimal input configuration from Chapter 3. Next, DCE-MR and diffusion weighted MR images are used for metastases detection within the liver mask.

The method developed in Chapter 4, is further developed to create a patient-specific lesion detection CNN in **Chapter 5**. Previously acquired MR scans of the patient are used to fine-tune the detection CNN. This fine-tuned CNN is tailored towards the specific features of the lesions and the surrounding healthy tissue of that patient, to improve the detection. The described patient-specific fine-tuning approach is also demonstrated on white matter hyperintensities in the brain.

In **Chapter 6**, a focal liver lesion classification method is described. Unlike the previous chapters, conventional machine learning is used. The classification method aims to differentiate between five lesion classes: adenomas, cysts, hemangiomas, HCCs and metastases. This chapter studies the impact of adding features from DCE-MR images and risk factors to the T2-weighted MR images, and will evaluate the proposed lesion classification method.

In **Chapter 7**, the main results of this thesis are summarized and future perspectives are discussed.

References

1. Kumar, P. & Clack, M. *Kumar and Clark's Clinical Medicine*. (Elsevier Saunders, 2005).

2. World Health Organization – International Agency for Cancer Research. *GLOBOCAN 2018, Cancer Incidence, Mortality and Prevalence Worldwide in 2018*. (2018).

3. Oniscu, G. C. & Parks, R. W. Benign conditions of the liver. *Surgery* **29**, 627–631 (2011).

4. Gomez, D. & Lobo, D. N. Malignant liver tumours. *Surg.* **29**, 632–639 (2011).

5. Kemeny, N. The management of resectable and unresectable liver metastases from colorectal cancer. *Curr. Opin. Oncol.* **22**, 364–373 (2010).

6. Sahani, D. V & Kalva, S. P. Imaging the liver. *Oncologist* **9**, 385–397 (2004).

7. Silva, A. C. *et al.* MR imaging of hypervascular liver masses: a review of current techniques. *Radiographics* **29**, 385–402 (2009).

8. Böttcher, J. *et al.* Detection and classification of different liver lesions : Comparison of Gd-EOB-DTPA-enhanced MRI versus multiphasic spiral CT in a clinical single centre investigation. *Eur. J. Radiol.* **82**, 1860–1869 (2013).

9. O'Connor, J. *et al.* Dynamic contrast-enhanced imaging techniques: CT and MRI. *Br. J. Radiol.* **84**, S112–S120 (2011).

10. Choyke, P. L., Dwyer, A. J. & Knopp, M. V. Functional tumor imaging with Dynamic Contrast- Enhanced Magnetic Resonance Imaging. *J. Magn. Reson. Imaging* **17**, 509–520 (2003).

11. Kenis, C. *et al.* Diagnosis of liver metastases : Can diffusion-weighted imaging be used as a stand alone sequence? *Eur. J. Radiol.* **81**, 1016–1023 (2012).

12. Albiin, N. MRI of focal liver lesions. *Curr. Med. Imaging Rev.* **8**, 107–116 (2012).

13. Hamy, V. *et al.* Respiratory motion correction in dynamic MRI using robust data decomposition registration - Application to DCE-MRI. *Med. Image Anal.* **18**, 301–313 (2014).

14. Pluim, J. P. W., Maintz, J. B. A. & Viergever, M. A. Mutual-information-based registration of medical images: a survey. *IEEE Trans. Med. Imaging* **22**, 986–1004 (2003).

15. Wells, W. M., Viola, P., Atsumi, H., Nakajima, S. & Kikinis, R. Multi-modal volume registration by maximization of mutual information. *Med. Image Anal.* **1**, 35–51 (1996).

16. Jain, A. K., Duin, R. P. . & Mao, J. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 4–37 (2000).

17. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image*

Anal. **42**, 60–88 (2017).

18. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

19. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. in *Proceedings of the 32nd International Conference on Machine Learning* **37**, (2015).

2

Evaluation of motion correction for clinical dynamic contrast enhanced MRI of the liver

Based on:
M.J.A. Jansen, H.J. Kuijf, W.B. Veldhuis, F.J. Wessels, M.S. van Leeuwen, and J.P.W. Pluim. "Evaluation of motion correction for clinical dynamic contrast enhanced MRI of the liver" *Physics in Medicine and Biology* 2017, 62(19):7556–68.

Abstract

Motion correction of 4D dynamic contrast enhanced MRI (DCE-MRI) series is required for diagnostic evaluation of liver lesions. The registration, however, is a challenging task, owing to rapid changes in image appearance. In this study, two different registration approaches are compared; a conventional pairwise method applying mutual information as metric and a groupwise method applying a principal component analysis based metric, introduced by Huizinga et al (2016)[1]. The pairwise method transforms the individual 3D images one by one to a reference image, whereas the groupwise registration method computes the metric on all the images simultaneously, exploiting the temporal information, and transforms all 3D images to a common space.

The performance of the two registration methods was evaluated using 70 clinical 4D DCE-MRI series with the focus on the liver. The evaluation was based on the smoothness of the time intensity curves in lesions, lesion volume change after deformation and the smoothness of spatial deformation. Furthermore, the visual quality of subtraction images (pre-contrast image subtracted from the post contrast images) before and after registration was rated by two observers.

Both registration methods improved the alignment of the DCE-MRI images in comparison to the non-corrected series. Furthermore, the groupwise method achieved better temporal alignment with smoother spatial deformations than the pairwise method. The quality of the subtraction images was graded satisfactory in 32% of the cases without registration and in 77% and 80% of the cases after pairwise and groupwise registration, respectively.

In conclusion, the groupwise registration method outperforms the pairwise registration method and achieves clinically satisfying results. Registration leads to improved subtraction images.

Introduction

Dynamic contrast enhanced MRI (DCE-MRI) is part of MRI examination of the liver and it involves scanning of the abdomen once before and several times after administration of a contrast agent. The images provide useful information about the microcirculation and tissue characteristics of liver lesions and parenchyma[2]. Subtraction images, resulting from the subtraction of the non-contrast image from the images after contrast administration, give additional insight into the contrast

uptake and washout of the lesion[3,4].

During the DCE-MRI acquisition the patient is asked to hold his/her breath several times to limit breathing motion, but the images can still suffer from inconsistencies because of varying breath hold depth, gasping for air[5], as well as cardiac and bowel movement. This motion between 3D acquisitions complicates the analysis of the 4D DCE-MRI series and the subtraction images can erroneously show enhancing lesions[3]. Motion correction improves the subtraction images by reducing motion artefacts[6] and supports more accurate analysis of lesions. An illustration of a subtraction image before and after motion correction is shown in Figure 1.

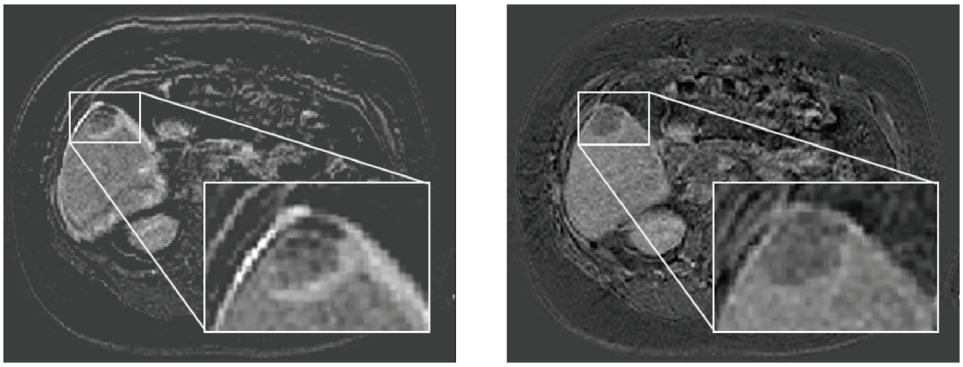


Figure 1 Differences between subtraction images before and after motion correction (left versus right) of an axial plane of the abdomen. The zoom-in of a lesion in the liver is inserted. Falsely enhancing edges of the liver and lesion are seen in the non-corrected image on the left. Motion correction avoids falsely enhancing edges of the liver and lesion.

Correcting for motion requires registration of the DCE-MRI series. A registration method should distinguish between intensity changes due to either the contrast agent and motion. Globally the motion is rigid, but locally it can be non-rigid and should be compensated in such way[7].

Over the years, several methods have been proposed to compensate for motion in DCE-MR series of the abdomen. Some approaches have used pairwise registration methods with mutual information (MI) as metric, where the individual images are registered one by one to a reference image[8–12]. These methods generally do not achieve the temporal smoothness needed for visual assessment, because of bias towards the choice of reference[1,13]. Groupwise registration methods for DCE-MRI have been proposed to achieve a better temporal smoothness. These methods distinguish motion from intensity changes due to contrast using pharmacokinetic models[14–16],

or data reduction techniques[5,17,18]. Pharmacokinetic models, describing the intensity changes due to contrast, have been used to drive registration by reducing the model fitting errors[14–16]. A potential downside is that pharmacokinetic modelling requires a high temporal resolution[19]. Another possible disadvantage is that a model could be corrupted by noise, or acquisition artefacts, leading to inappropriate model fits and thus misregistration[1,17]. Some registration methods apply data reduction techniques, such as principle component analysis (PCA) or independent component analysis (ICA), to separate motion from contrast induced intensity changes. These approaches iteratively create synthetic images based on the results of data reduction and register towards the synthetic images, as in Hamy et al (2014), Melbourne et al (2007), and Wollny et al (2012)[5,17,18]. All these proposed approaches are not intrinsically groupwise approaches, since they register towards synthetic images based on data reduction or pharmacokinetic models, which could cause a bias towards these images. A true groupwise registration method would apply the cost function on the 4D image and aligns the images without the use of a reference image. Such a method is proposed by Huizinga et al (2016)[1] and this PCA-based method will be validated on a clinical data set with a low temporal resolution (16 time points in 5 minutes).

In this chapter, clinical data sets were used to compare a conventional pairwise registration approach, using MI as metric, with the PCA-based groupwise registration method[1]. The two methods were compared based on quantitative evaluation methods as well as on the visual quality of the subtraction images, providing a good evaluation for clinical application.

Materials and Methods

Data

Data sets of 70 patients from the UMC Utrecht were included in the study. All patients had a 4D DCE-MRI scan with the clinical focus on the liver. The DCE-MRI images were acquired in six breath holds with 1 to 5 3D images per breath hold. The DCE-MRI series was acquired on a 1.5 T scanner (Phillips) using a clinical protocol with the following parameters: TE: 2.143 ms; TR: 4.524 ms; flip angle: 10 degrees. After acquiring the first image of the series, gadobutrol (0.1 ml/kg Gadovist of 1.0 mmol/ml at 1 ml/s) or gadoxetate disodium (0.1 ml/kg Primovist of 0.25 mmol/ml at 1 ml/s) was administered at once, followed by 25 ml saline solution at 1 ml/s. In total 16 3D images per patient were acquired with 90-105 slices and a matrix of 256×256. Voxel size was 1.543 mm ×1.543 mm × 2 mm. Further details can be found in Schalkx et al (2014)[20].

One observer annotated one lesion in one slice of the non-contrast image for 49 data sets. The other data sets did not contain lesions. The 2D annotations were expanded to 3D using Otsu-thresholding, with thresholds based on the annotated slice, and hole-filling. This resulted in a rough lesion segmentation.

Registration methods

All the data sets were registered using two different registration approaches. One is a conventional pairwise registration method, using MI as metric. The other is a groupwise registration method, using a PCA-based dissimilarity metric. Both methods are available in the open-source software package elastix[21], enabling a direct comparison.*

Pairwise registration

Mutual information (MI) was chosen as cost function for the pairwise registration method. MI is robust for intensity differences, due to the contrast inflow, between the moving and the reference image. The first, non-contrast image was used as the reference image to which all other images in the DCE-MRI series were registered, because this image is subtracted from the contrast images to yield the subtraction images. The registration was done in three resolutions using a smoothing image pyramid schedule with a downsampling factor of 4, 2 and 1 in each dimension for the three successive resolutions. The adaptive stochastic gradient descent optimizer was employed for 500 iterations each resolution, with 2048 new samples every iteration. The mutual information between the two images was calculated in 32 histogram bins. A B-Spline transformation with a final grid spacing of 16 mm was applied.

Groupwise registration

The groupwise registration method assumes that the intensity changes due to contrast can be described as a low dimensional signal model without prior knowledge of this model[1,17,18]. The method separates the intensity changes due to contrast from motion and registers all images simultaneously to a common space by minimizing a cost function based on PCA. The cost function is calculated as follows: the 4D DCE-MRI series with V images and N voxels per 3D image is reshaped to an N × V matrix M. The correlation matrix C of this matrix is defined as:

$C = \frac{1}{N-1} S^{-1}(M - \bar{M})^T(M - \bar{M})S^{-1}$,where $\bar{M}$ is a columnwise mean of matrix M, and

* The parameter files used, are available at elastix.bigr.nl/wiki/index.php/Par0047

S is diagonal matrix with the standard deviations of each column in *M*. The eigenvalues of *C* are obtained by principle component analysis. The cost function is defined as the sum of the inversely weighted eigenvalues: $D_{PCA} = \sum_{j=1}^{V} j\lambda_j$, with j the rank of the eigenvalues. In order to maximize the highest eigenvalues, and impose more alignment in the DCE-MRI series, this cost function is minimized[1].

The registration was done in four resolutions using a smoothing image pyramid schedule with a downsampling factor of 8, 4, 2 and 1 in each dimension, except for the time dimension, which was not down sampled. The adaptive stochastic gradient descent optimizer was employed for 500 iterations each resolution, with 2048 new samples every iteration. A B-Spline stack transform with a final grid spacing of 16 mm was applied.

A fourth resolution, with downsampling factor 8, was effective for the groupwise registration with multiple images, but did visually not result in better registration in the pairwise registration approach, where only one image was deformed.

Experiments

Two types of experiments were performed; quantitative and qualitative evaluations. The quantitative evaluation aims to assess the accuracy of both registration methods and the qualitative evaluation aims to assess the clinical quality and reliability of the subtraction images after registration. The two registration methods were compared and, when possible, the original non-corrected images were included in the evaluation.

Quantitative evaluation

Assessment of registration accuracy is complicated, because the liver does not have many distinguishable landmarks[1] and manual placement of these landmarks is a difficult task, prone to errors[22]. Therefore the evaluation was based on elements important in the clinical practice.

The ability of the registration methods to align the images of the DCE-MRI series was assessed on (1) temporal intensity smoothness, (2) lesion volume change and (3) spatial smoothness. Criteria 1 and 2 were calculated on the lesions, since these are the clinical regions of interest and were therefore evaluated on 49 datasets; criterion 3 was applied to all 70 data sets. The masks were obtained for the non-contrast image, which is the reference image for pairwise registration. For the groupwise registered calculations, the masks were transformed towards the common space of the groupwise

registered series, to compare the same voxels in non-corrected, pairwise registered and groupwise registered DCE-MRI series. For both the pairwise and the groupwise registration, the transformations are defined for each 3D image separately.

Temporal intensity smoothness

The time intensity curve of contrast in a lesion should be smooth. A time intensity curve can be disrupted by motion caused by inconsistent breath hold, see Figure 2. For each voxel in the lesion mask, the standard deviation (SD) of the second derivative of the time intensity curve was calculated. The mean of all these standard deviations is an indication of the temporal intensity smoothness in the lesion. This temporal intensity smoothness was calculated in the 49 subjects with a lesion, for the non-corrected series, the pairwise registered and the groupwise registered series. Some lesions have different time intensity curves depending on the presence and rate of contrast uptake. The outcomes were thus compared on patient level with the Wilcoxon signed rank test.

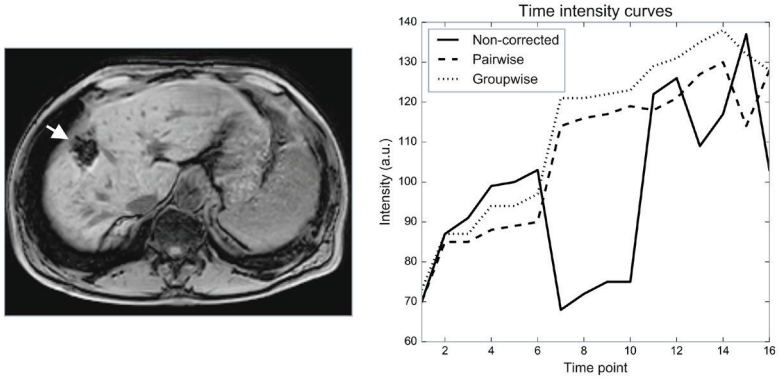


Figure 2 Time intensity curves of a lesion showing severe motion during the acquisition of the DCE-MRI series. On the left the axial view of the non-contrast image is shown. The arrow indicates the lesion. On the right the mean intensity curves of that lesion are shown for the non-corrected DCE-MRI series and the pairwise and groupwise registered DCE-MRI series. The consequences of an inconsistent breath hold are visible in time points 7 to 10 of the non-corrected series. Registration corrects for this displacement.

Lesion volume change

Improper registration could lead to changes in lesion volume[6]. Volume expansion or shrinkage for each voxel after deformation is defined by the Jacobian determinant of the deformation. The Jacobian determinant is defined on the reference (or common space) image voxels towards the moving image. Hence, only the reference lesion mask is needed. The geometric mean of the Jacobian determinant in the lesion was

calculated for each image to identify volume changes induced by transformation. A mean of 1.00 indicates no volume change. The geometric mean Jacobian determinant was calculated for all transformed images, including the first, non-contrast image in the groupwise registration method. A t-test was performed between the outcomes of the pairwise and the groupwise registration methods.

Spatial smoothness

The SD of the Jacobian determinant of the deformation is a measure for the smoothness of the transformation in the spatial domain. A smaller SD implies a smoother spatial transformation. The SD of the Jacobian determinant in the liver region was obtained for each transformed 3D image, including the first non-contrast image in the groupwise registration. A t-test was performed between the outcomes of the pairwise and the groupwise registration methods.

Qualitative evaluation

Registration quality was evaluated through visual inspection of subtraction images by two observers. The subtraction images were obtained by subtracting the first, non-contrast image of the time series from the remaining images of the time series, which are contrast enhanced images. The subtraction images of the non-corrected, the pairwise registered and the groupwise registered DCE-MRI series (n=3×70) were shown in a random order to two observers, radiologists with clinical experience of more than 10 years, who were unaware of the registration status of the images. The observers graded the subtraction images independently on a 5-point scale: 1) poor quality; useless and unrealistic subtraction image, 2) mediocre quality; difficult to assess whether the contrast enhancement is real or a result of registration errors, 3) acceptable quality; but a few erroneous enhancing areas as a result of registration errors, 4) good quality; assessment of contrast enhancement is easy, but some small unimportant erroneous enhancement visible, 5) excellent quality; assessment of contrast enhancement is easy and no erroneous enhancement visible. A grade of 4 or 5 means that the quality is satisfactory for the observers. The focus was mainly on the correctness of the subtraction image, but the alignment between successive images was also considered. The non-corrected, non-subtracted DCE-MRI series were always supplied as a reference for the enhancement and to assess the quality of the subtraction images in relation to the quality of the non-corrected, non-subtracted images.

The grades by the two observers were averaged for evaluation, unless the grades

differed more than one point, in which case consensus was reached by the two observers. A difference of one point between the grades was acceptable, because the distinction between two adjacent grades is subjective.

The Wilcoxon signed rank test was used to test for significant differences between the visual quality of the subtraction images of the non-corrected, the pairwise registered and the groupwise registered DCE-MRI images. The inter-observer agreement was calculated using Spearman's rank correlation coefficient.

Results

Quantitative evaluation

Temporal intensity smoothness

The median (interquartile range (IQR)) of the mean SD of the second derivative of the time intensity curve is 22.0 (11.4) for the non-corrected images, 22.6 (9.2) for the pairwise registered images and 16.6 (7.6) for the groupwise registered images. A lower mean SD means smoother intensity transitions between successive 3D images. Figure 3 shows the distribution of the mean SD of the subjects. The Wilcoxon signed rank test shows that the groupwise registration differs significantly from the non-corrected images and the pairwise registered images ($p = 0.000$ for both). The difference found between the non-corrected images and the pairwise images is less distinct ($p = 0.046$).

Lesion volume change

The volume change of the lesions after pairwise and groupwise transformation is plotted in Figure 4. In the boxplot, the geometric mean of the Jacobian determinant of the lesion is shown for each subject and each time point. The median geometric mean (IQR) is 0.99 (0.15) for pairwise registration and 1.00 (0.06) for groupwise registration. A t-test showed that there is no significant difference between the two groups ($p=0.787$). The 24 outliers of the pairwise transformation originated from six subjects, two of which had seven or more time points in which the lesion volume change was defined as an outlier (7 and 10). In the other subjects only three outliers or fewer were found. For the groupwise results the 38 outliers originated from 16 subjects, two of which had seven or more outliers (7 and 10). In the other subjects only three time points or fewer had a lesion volume change defined as an outlier.

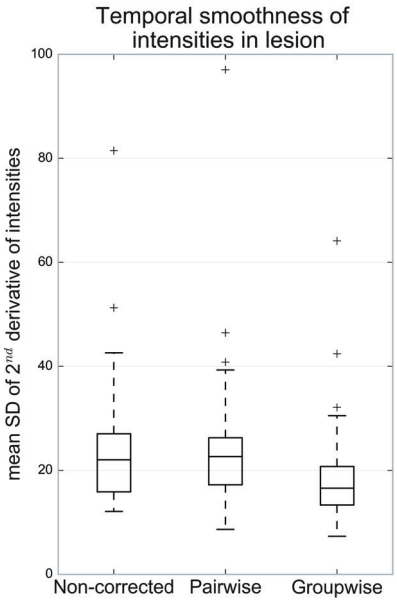


Figure 3 Smoothness of time intensity curves in lesions before and after pairwise and groupwise registration.

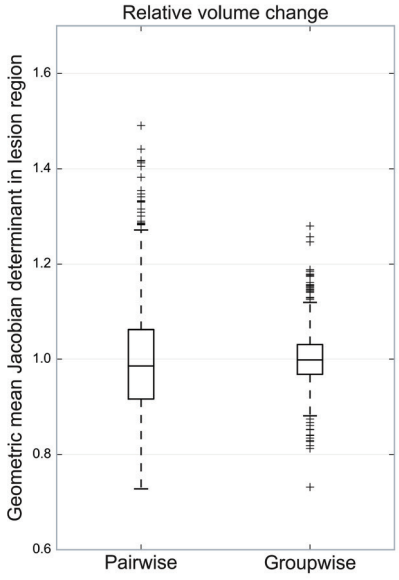


Figure 4 Relative volume changes of lesion regions per transformed 3D image, for pairwise and for groupwise registration.

Spatial smoothness

The SD of the Jacobian determinant in the liver region is calculated in each time point of the 70 data sets. The median (IQR) SD of the Jacobian determinant for the pairwise deformation is 0.115 (0.047) and 0.044 (0.020) for the groupwise deformation, see Figure 5. A t-test showed a significant difference between the pairwise and groupwise results (p = 0.000).

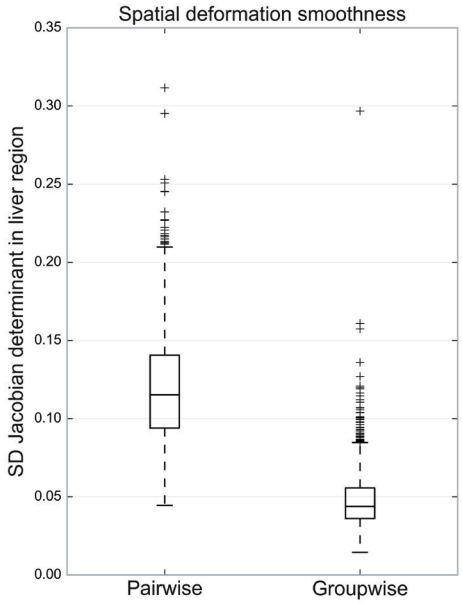


Figure 5 Spatial deformation smoothness of liver regions per transformed 3D image.

Qualitative evaluation

The subtraction images of the original, pairwise and groupwise registered images received identical scores from the two observers in 33% of cases, whereas a difference of one, two and three points was seen between the observers in 46%, 18% and 3% of cases, respectively. A moderate inter-observer agreement was observed with a Spearman's rank correlation coefficient of 0.444 (p = 0.000).

For cases that differed by more than one point, consensus was reached by the observers. The mean grade after averaging and consensus (SD) was 2.9 (0.7), 3.8 (0.9) and 4.0 (0.8) for non-corrected, pairwise and groupwise registration respectively. The

Wilcoxon signed rank test showed a significant difference between non-corrected and pairwise registered images (p = 0.000), between non-corrected and groupwise registered images (p = 0.000) and between pairwise and groupwise registered images (p = 0.035). Figure 6 shows the frequency of the assigned grades.

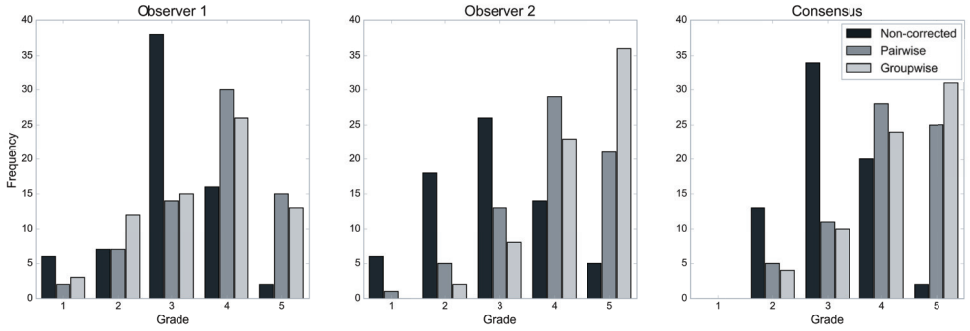


Figure 6 Histogram plots of grades by observer 1, observer 2 and the consensus grades.

The original non-corrected DCE-MRI series was judged satisfactory (grade 4 or 5) in 32% of the images. The pairwise registered images have satisfactory results in 77% of the images and the groupwise registered images in 80% of the images.

A positive difference of more than one point between the consensus grades of the non-corrected and registered subtraction images was observed in 16 subjects (23%) for the pairwise method and in 27 subjects (39%) for the groupwise method, see Figure 7. A negative difference of more than one point between the consensus grades of the non-corrected and registered subtraction images was observed in 1 subject (1.4%) for the pairwise method and not present for the groupwise method.

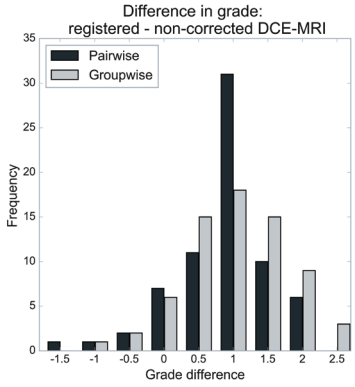


Figure 7 Difference between original and registered image grades for consensus grading.

Discussion

In this study, two different registration approaches were evaluated for motion correction of clinical DCE-MRI data. The evaluation was directed both at quantitative comparison of method results and at clinical application of motion corrected images, in the form of subtraction of non-contrast from contrast images.

Quantitative evaluation

The lack of a ground truth is alleviated by performing multiple experiments. The experiments evaluated the temporal intensity smoothness, lesion volume change and spatial smoothness after deformation.

Remarkably, the pairwise registration method gave slightly worse results than the non-corrected images in the temporal intensity smoothness evaluation. In 29 out of 49 subjects the average SD of the second derivative of the temporal intensity curve was higher for the pairwise registered images than for the non-corrected images, but the differences are small. Visual inspection showed that successive images of the non-corrected series within the same breath hold are often already aligned, but not necessarily with the non-contrast image. Pairwise registration improves the alignment with the non-contrast image at some cost of alignment between successive images. Although the temporal intensity smoothness of the pairwise registered images is slightly worse than that of the non-corrected images, the evaluation of the visual quality of the subtraction images shows that this did not result in poorer quality.

When it comes to lesion volume change, no significant difference is found between the two registration methods, however the distribution is wider for the pairwise registration than for the groupwise registration. The groupwise method is better able to preserve lesion volume than the pairwise method.

The third experiment evaluated the spatial smoothness after deformation. The median SD of the Jacobian determinant of the pairwise deformation was higher than for the groupwise deformation. The same pattern was seen by Huizinga et al (2016)[1] on different data sets with a higher temporal resolution.

The difference between the pairwise and groupwise registration results can mostly be explained by the bias towards the reference image. The groupwise registration method eliminates this bias by simultaneous registration to a common space and applying the metric to all images at once, instead of separately registering toward a reference

image. Although the groupwise method does not enforce temporal smoothness, like Metz et al (2011)[13], the use of temporal information accomplishes better temporal alignment of the images.

Apart from the differences in performance, the methods also differ in processing time and computational memory. The total run time of an entire DCE-MR series is ~11 minutes for the pairwise registration and ~40 minutes for the groupwise registration approach on a standard workstation. The groupwise registration approach requires more memory to load all the 16 images at once, around 5 GB, which is far more than the 0.3 GB needed for the pairwise approach that only loads two images at once. Both methods are implemented in elastix, which is freely available at elastix.isi.uu.nl.

Qualitative evaluation

The visual quality of the subtraction images before and after registration were evaluated by two observers. The evaluation was done in a subjective manner, hence the moderate inter-observer agreement.

Both observers graded the registered images higher than the non-corrected images. Overall, observer 1 had a slight preference for the subtraction images after pairwise registration, while observer 2 preferred the subtraction images after groupwise registration. There are two main reasons for the moderate inter-observer agreement. First, despite the clear instructions in section 2.3.2, both observers had a slightly different personal interpretation of the grading scale, something that can hardly be avoided in visual grading. Second, each observer might have focussed on different time points in the temporal dimension of the subtraction images. In the workflow, observers quickly glance over the full temporal dimension, after which they focus on the lesion or structure of interest in only a few time points. For these reasons, we opted for a consensus reading in case of more than one point difference between the observers. After consensus and averaging, the groupwise approach obtained higher grades than the pairwise approach.

Improvement of the quality of the subtraction images due to registration was also observed in a study by Sundarakumar et al (2015)[6]. In this study, the two observers had a higher inter-observer agreement than observed in our work. This study compared the ability of a pairwise registration approach to align the non-enhanced, arterial and delayed phase with the portal-venous phase. The quality assessment was done for each phase individually, explaining the difference in inter-observer agreement. The preference for groupwise registered images over non-corrected or pairwise registered

images was also shown by Melbourne et al (2007)[17] for a different registration method and different data. However, here the preference between two (registered) images was not given by clinical experts, but by the researchers.

Because of the satisfying quality of the subtraction images created by the groupwise registration, this method has been implemented in the daily clinical routine for abdominal DCE-MR images.

Conclusions

Registration of the DCE-MRI series of the liver improves the alignment of the images by correcting for the motion introduced by inconsistent breath hold, bowel and cardiac movements. It also provides the opportunity for image analysis in the clinical workstation, like contrast kinetic curves and signal intensity characterization. This study shows that the groupwise registration using a PCA-based dissimilarity metric on DCE MRI of the liver achieves better temporal alignment with smoother spatial deformations than pairwise registration with a mutual information metric. Besides this, motion correction increased the amount of images with a satisfying quality of the subtraction images from 32% to 77% and 80% of the cases after pairwise and groupwise registration, respectively.

References

1. Huizinga, W. *et al.* PCA-based groupwise image registration for quantitative MRI. *Med. Image Anal.* **29**, 65–78 (2016).

2. Choyke, P. L., Dwyer, A. J. & Knopp, M. V. Functional tumor imaging with Dynamic Contrast- Enhanced Magnetic Resonance Imaging. *J. Magn. Reson. Imaging* **17**, 509–520 (2003).

3. Yu, J. & Rofsky, N. M. Dynamic Subtraction MR Imaging of the Liver : Advantages and Pitfalls. *AJR* **180**, 1351–1357 (2003).

4. Yu, J., Kim, Y. H. & Rofsky, N. M. Dynamic Subtraction Magnetic Resonance Imaging of Cirrhotic Liver : Assessment of High Signal Intensity Lesions on Nonenhanced T1-weighted Images. *J Comput Assist Tomogr* **29**, 51–58 (2005).

5. Wollny, G., Kellman, P., Santos, A. & Ledesma-Carbayo, M. J. Automatic motion compensation of free breathing acquired myocardial perfusion data by using independent component analysis. *Med. Image Anal.* **16**, 1015–1028 (2012).

6. Sundarakumar, D. K., Wilson, G. J., Osman, S. F., Zaidi, S. F. & Maki, J. H. Evaluation of image registration in subtracted 3D dynamic contrast-enhanced MRI of treated hepatocellular carcinoma. *Am. J. Roentgenol.* **204**, 287–296 (2015).

7. Melbourne, A., Atkinson, D. & Hawkes, D. Influence of Organ Motion and Contrast. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 948–955 (2008).

8. Rueckert, D. *et al.* Nonrigid Registration Using Free-Form Deformations : Application to Breast MR Images. *IEEE Trans. Med. Imaging* **18**, 712–721 (1999).

9. Rohlfing, T., Maurer, C. R., Bluemke, D. A. & Jacobs, M. A. Volume-Preserving Nonrigid Registration of MR Breast Images Using Free-Form Deformation With an Incompressibility Constraint. **22**, 730–741 (2003).

10. Wells, W. M., Viola, P., Atsumi, H., Nakajima, S. & Kikinis, R. Multi-modal volume registration by maximization of mutual information. *Med. Image Anal.* **1**, 35–51 (1996).

11. Hodneland, E. *et al.* Segmentation-Driven Image Registration- Application to 4D DCE-MRI Recordings of the Moving Kidneys. *IEEE Trans. Image Process.* **23**, 2392–2404 (2014).

12. Zöllner, F. G. *et al.* Assessment of 3D DCE-MRI of the kidneys using non-rigid image registration and segmentation of voxel time courses. *Comput. Med. Imaging Graph.* **33**, 171–181 (2009).

13. Metz, C. T., Klein, S., Schaap, M., van Walsum, T. & Niessen, W. J. Nonrigid registration of dynamic medical imaging data using nD+t B-splines and a groupwise optimization approach. *Med. Image Anal.* **15**, 238–249 (2011).

14. Buonaccorsi, G. A. *et al.* Tracer kinetic model-driven registration for dynamic contrast-enhanced MRI time-series data. *Magn. Reson. Med.* **58**, 1010–1019 (2007).

15. Hayton, P., Brady, M., Tarassenko, L. & Moore, N. Analysis of dynamic MR breast images using a model of contrast enhancement. *Med. Image Anal.* **1**, 207–224 (1997).

16. Bhushan, M. *et al.* Motion correction and parameter estimation in DCE MRI sequences : Application to colorectal cancer. *Lect. Notes Comput. Sci. MICCAI 2011* 476–483 (2011).

17. Melbourne, A. *et al.* Registration of dynamic contrast-enhanced MRI using a progressive principal component registration (PPCR). *Phys. Med. Biol.* **52**, 5147–5156 (2007).

18. Hamy, V. *et al.* Respiratory motion correction in dynamic MRI using robust data decomposition registration - Application to DCE-MRI. *Med. Image Anal.* **18**, 301–313 (2014).

19. Henderson, E., Rutt, B. K. & Lee, T.-Y. Temporal sampling requirements for the tracer kinetics modeling of breast disease. *Magn. Reson. Imaging* **16**, 1057–1073 (1998).

20. Schalkx, H. J. *et al.* Liver perfusion in dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI): comparison of enhancement in Gd-BT-DO3A and Gd-EOB-DTPA in normal liver parenchyma. *Eur. Radiol.* **24**, 2146–2156 (2014).

21. Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. W. elastix: A toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* **29**, 196–205 (2010).

22. Murphy, K. *et al.* Semi-automatic construction of reference standards for evaluation of image registration. *Med. Image Anal.* **15**, 71–84 (2011).

3

Optimal input configuration of dynamic contrast enhanced MRI in convolutional neural networks for liver segmentation

Based on:
M.J.A. Jansen, H.J. Kuijf, and J.P.W. Pluim, "Optimal input configuration of dynamic contrast enhanced MRI in convolutional neural networks for liver segmentation", in: *SPIE Medical Imaging* 2019, 10949, p. 1094941V.

Abstract

Most MRI liver segmentation methods use a structural 3D scan as input, such as a T1 or T2 weighted scan. Segmentation performance may be improved by utilizing both structural and functional information, as contained in dynamic contrast enhanced (DCE) MR series. Dynamic information can be incorporated in a segmentation method based on convolutional neural networks in a number of ways. In this study, the optimal input configuration of DCE MR images for convolutional neural networks (CNNs) is studied.

The performance of three different input configurations for CNNs is studied for a liver segmentation task. The three configurations are I) one phase image of the DCE-MR series as input image; II) the separate phases of the DCE-MR as input images; and III) the separate phases of the DCE-MR as channels of one input image. The three input configurations are fed into a dilated fully convolutional network and into a small U-net. The CNNs were trained using 19 annotated DCE-MR series and tested on another 19 annotated DCE-MR series. The performance of the three input configurations for both networks is evaluated against manual annotations.

The results show that both neural networks perform better when the separate phases of the DCE-MR series are used as channels of an input image in comparison to one phase as input image or the separate phases as input images. No significant difference between the performances of the two network architectures was found for the separate phases as channels of an input image.

Introduction

Automatic liver segmentation in abdominal MR images is an important step in automatic detection and characterization of focal liver lesions. Most (semi-)automatic MR liver segmentation methods are designed to process an anatomical 3D image as input instead of the dynamic contrast enhanced (DCE) MR images.[1–4] These DCE-MR images hold valuable information about liver shape and function in the different contrast phases. Using the rich information of the DCE-MR images in convolutional neural networks (CNNs) might improve the segmentation performance.

However, it is unclear how dynamic MR series are best incorporated in CNNs. There are several ways to process the series in a CNN. The conventional way is to use one of the 3D phase images of the DCE-MR series as input of a CNN, but this neglects the dynamic nature of the DCE-MR series. To include the complete series of phases,

rather than a single one, we investigate two different approaches. One is to process each phase separately in the CNN and merge the feature maps of the different phases at the end of the CNN. The second approach is to use the phases as different channels of the input image, like an RGB image. Each of the latter two configurations contains identical input information, but it is processed differently, which might result in different outcomes.

In this chapter, we investigate the performance of two commonly used CNN architectures using the three different input configurations as described above. Both the U-net architecture[5] and a dilated fully convolutional network (FCN) architecture[6] were selected, based on their good performance in medical image segmentation tasks[7]. The two architectures are compared in order to study the effect of the input configuration regardless of the network architecture.

Methods

Data

The DCE-MR series were acquired in six breath holds with one to five 3D images per breath hold. The DCE-MRI series was acquired on a 1.5 T MRI scanner (Philips, Best, The Netherlands) using a clinical protocol with the following parameters: TE: 2.143 ms; TR: 4.524 ms; flip angle: 10 degrees. After acquiring the first image, gadobutrol (0.1 ml/kg Gadovist of 1.0 mmol/ml at 1 ml/s) was administered at once, followed by 25 ml of a saline solution at 1 ml/s. In total, 16 3D images per patient were acquired with 100 slices and matrix sizes of 256 × 256. Voxel size was 1.543 mm × 1.543 mm × 2 mm.

The DCE-MR series of thirty-eight patients from the University Medical Center Utrecht, The Netherlands were used in this work. The data sets were randomly divided in a training set, validation set and test set. Sixteen data sets were used for training, three for validation, and nineteen for testing. The DCE-MR series were corrected for motion using the PCA-based groupwise registration algorithm, introduced in Chapter 2. A zero-mean-unit-variance rescaling was applied to all intensity values between the 0^{th} and 99.8^{th} percentile of the intensity histogram for intensity normalization. In this study we assumed that the 99.8^{th} percentile intensity corresponds to the contrast agent peak in the aorta. Intensities higher than the 99.8^{th} percentile are most probably noise and were assigned the same intensity value as the 99.8^{th} percentile intensity.

The DCE-MR images were combined per breath hold by averaging over the fourth

dimension, resulting in six contrast phases of DCE-MRI. The series start with the pre-contrast image. This is followed by the other phases: the early arterial phase, the late arterial phase, the hepatic/portal-venous phase, the late portal-venous/equilibrium phase and the late equilibrium phase.

The liver was manually annotated by two observers. The training and test sets were annotated by the first observer. The test set was annotated twice by the first observer with at least one week in between the two annotations and once by a second observer. A radiologist with more than 10 years of experience in liver MR analysis verified all manual segmentations and provided corrections where needed. Most MR data sets had at least one liver lesion, except for three data sets in the training set, one in the validation set and four in the test set. The liver lesions were included in the liver segmentations. Figure 1 shows two examples of DCE-MRI scans with a green contour around the liver and a red contour around the liver lesions.

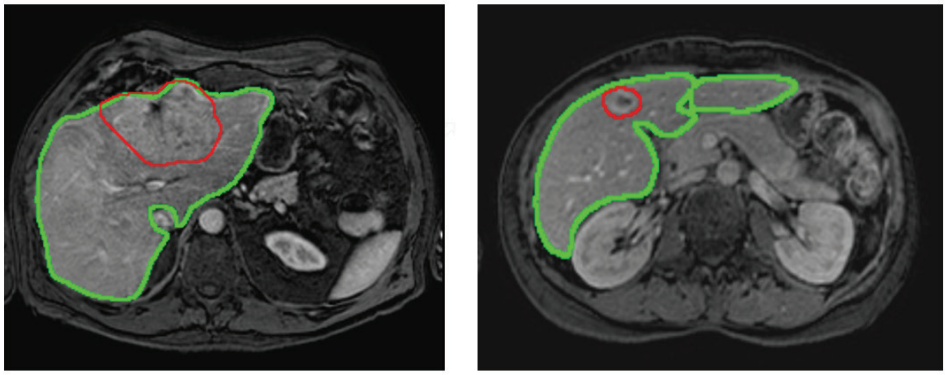


Figure 1 Two examples of the late arterial phase of the DCE-MRI scans. The livers are contoured in green. The lesions are contoured in red.

Inputs

Three input possibilities of DCE-MRI for CNNs were evaluated:

I. One phase of the DCE-MRI with the best visual contrast between the liver and other abdominal organs (late arterial phase);

II. All six phases of the DCE-MRI processed by the network one by one and merged at the second last layer of the network;

III. All six phases of the DCE-MRI as channels of the input image.

Network architectures

Dilated fully convolutional network The dilated FCN[6] consists of 7 layers with each a 3×3 convolution and 32 kernels. The 3rd layer has a dilation of 2, the 4th layer a dilation of 4, the 5th layer of 8 and the 6th layer of 16. The final 8th layer has a 1×1 convolution. This gives a receptive field of 67×67 pixels.

Modified U-net The second network is smaller than the conventional U-net[5], it only has four stages with three skip connections. We choose a smaller network because our images are only 256×256 voxels and an extra layer of downsampling would result in small feature maps. Also the number of kernels was reduced to 16 in the upper layers and to 128 kernels in the bottom layers.

For the input approach in which the phase images are used as separate input images (configuration II), the feature maps of the six phases are concatenated and an extra 1×1 convolution layer is added before the final 1×1 convolution layer in both networks, to combine the feature maps. Figure 2 gives an overview of both network architectures.

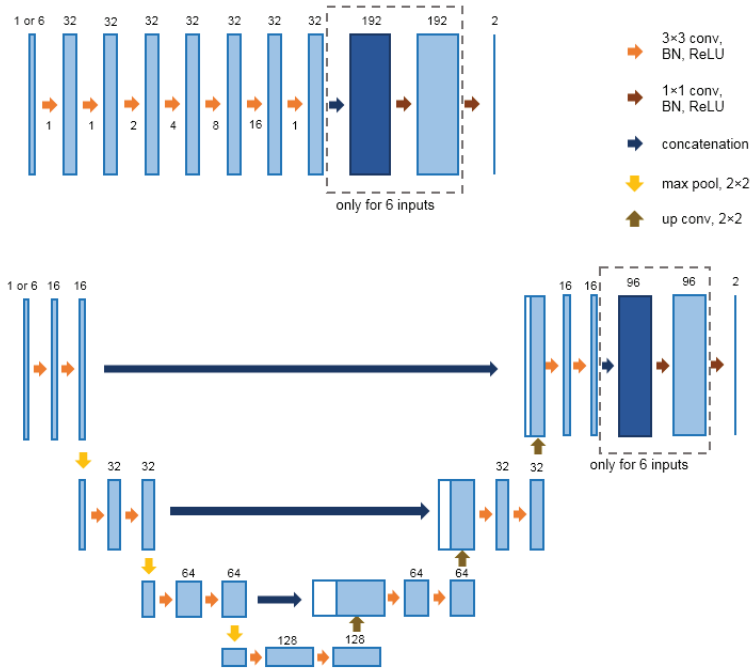


Figure 2 The dilated fully convolutional network and the modified U-net. The dilation rate of the convolution in the dilated FCN is given below the orange arrow, no digit means no dilation. The number of kernels or channels are given above the light blue rectangles. The number above the dark blue rectangle is the size of the merged feature maps. The white boxes in the modified U-net are the copied feature maps.

The loss is calculated by a similarity metric: $\frac{2*X \cap Y+s}{X^2+Y^2+s}$, where X is the predicted segmentation, Y is the ground truth mask and s is a small number to prevent dividing by zero and set to 1e-5[8]. ReLU activation and batch normalization are used in all the convolutional layers, except for the final layer which has a softmax activation. Glorot uniform initialization[9] and Adam optimization with a learning rate of 0.001 are used. The networks were trained for 500,000 iterations, with mini batches of one slice per mini batch. No data augmentation was applied.

To limit the parameters and training time for this feasibility study, both networks were trained on 2D slices, but the concept can easily be expanded to 3D.

Post-processing

The post-processing consists of setting a threshold of 0.50 on the probability outcome of the networks. After that 3D hole filling is performed, to fill the holes caused by liver lesions. The largest connected component is selected as the final segmentation.

Experiments

The performance of the three input configurations is evaluated based on the Dice similarity coefficient (DSC) and the modified Hausdorff distance (HD) based on the 95th percentile. The first set of liver annotations of the first observer is considered the ground truth.

DSC: $\frac{2\,X \cap Y}{X+Y}$, where X is the predicted segmentation and Y the ground truth mask.

HD at 95th percentile: $\max\left(h_{95}(X,Y), h_{95}(Y,X)\right)$, with $h_{95}(X,Y) = K_{x\in X}^{95^{th}}\, min_{y\in Y}\|y-x\|$. Where K^{95tl} is the 95th ranked minimum of the minimum Euclidean distances between all points of y and x.

These two metrics are computed on the segmentation results for each input configuration and network architecture. The DSC and the loss function are both similarity metrics and thus the outcomes of the CNNs are optimized for the DSC. However, the DSC can be used to compare the segmentation results of different network architectures and input configurations, since the loss function is always the same.

Inter- and intra-observer agreement

The inter- and intra-observer agreements for manual segmentation of the liver in the DCE-MR series in the test set were obtained to compare the performance of the network segmentations with human annotations. The same metrics as mentioned above are calculated for the inter- and intra-observer agreement.

Statistics

The paired Student's t-test was calculated for the DSC results between the three input configurations, for each of the two networks. Additionally, it was computed between the DSC results of the two network architectures with the six channels image inputs (configuration III). The paired Student's t-test was computed on the DSC results of configuration III and the DSC results of the inter-observer agreement, to test for a significant difference. The same was done for the intra-observer agreement DSC results.

The Wilcoxon signed rank test was calculated for the HD results between the three input configurations, for each of the two networks. Additionally, it was calculated between the HD results of the two network architectures for configuration III. To test the difference between the configuration III results and the inter-observer agreement results, the Wilcoxon signed rank test was computed on the HD results. The same was done for the intra-observer agreement HD results.

Results

The DSC and HD values of the liver segmentations of the three input configurations for the two CNNs are given in Figure 3.

The median [interquartile range (IQR)] inter-observer agreement is 0.943 [0.933 - 0.954] for the DSC and 5.56 mm [4.29 - 6.75 mm] for the HD. The median [interquartile range (IQR)] intra-observer agreement is 0.959 [0.956 - 0.966] for the DSC and 3.09 mm [2.53 - 3.09 mm] for the HD.

Paired Student's t-tests show a significant difference ($p \geq 0.05$) between the DSC of configuration I vs III for the dilated FCN, between configuration II and III for both the dilated FCN and the U-net, and between configuration I vs II for the U-net.

A significant differences was seen for the HD between configuration II and III for

the U-net. No significant difference was seen between the dilated FCN and the U-net architectures when the third configuration is used.

The DSC and HD of the computed liver segmentations differ significantly from the DSC and HD of the inter- and intra-observer results in the two metrics for both networks with the best input; configuration III.

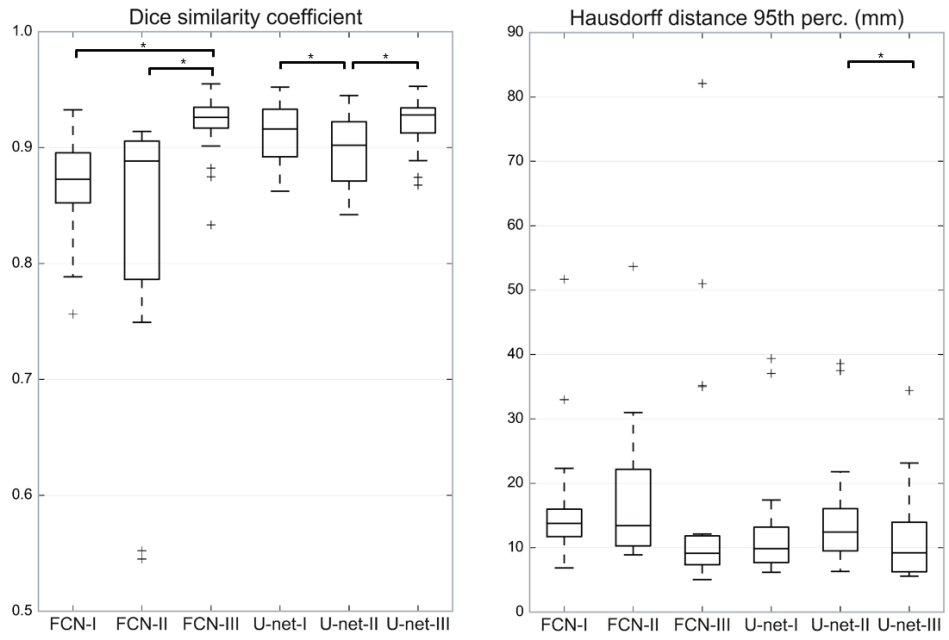


Figure 3 Boxplots of the dice similarity coefficient and modified Hausdorff distance (mm) for the dilated FCN and the modified U-net using the three input configurations. The numbers I, II, and III correspond to the three input configurations: one phase image, six phase images, and six phases as channels input image, respectively. An asterisk indicates a significant difference ($p \geq 0.05$) between two input configurations.

Examples of the segmentation results of one slice for the three input configurations and the two neural networks are given in Figure 4. It shows a typical result, where configuration III shows the best result for both network architectures. Segmentation results of a liver with a large lesion are given in Figure 5. A great part of the lesion is included in the segmentations. Configuration II has the worst results with regard to the lesion for both networks.

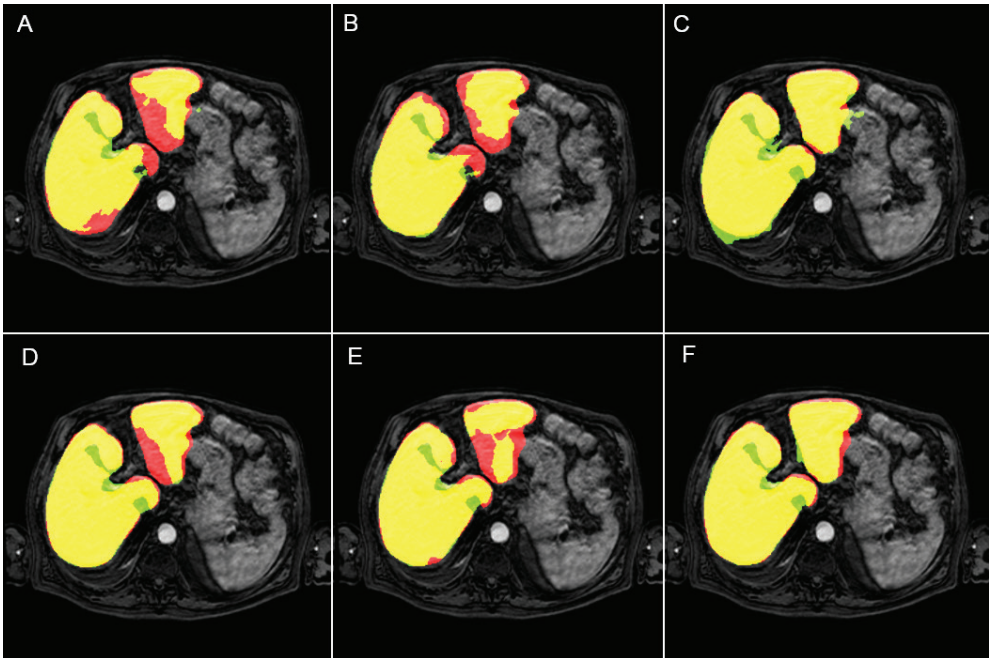


Figure 4 Segmentation results after post-processing. The first row represents the dilated FCN and the second row the modified U-net. From left to right the columns represent configurations I, II, and III. Yellow is true positive, red is false negative, and green is false positive.

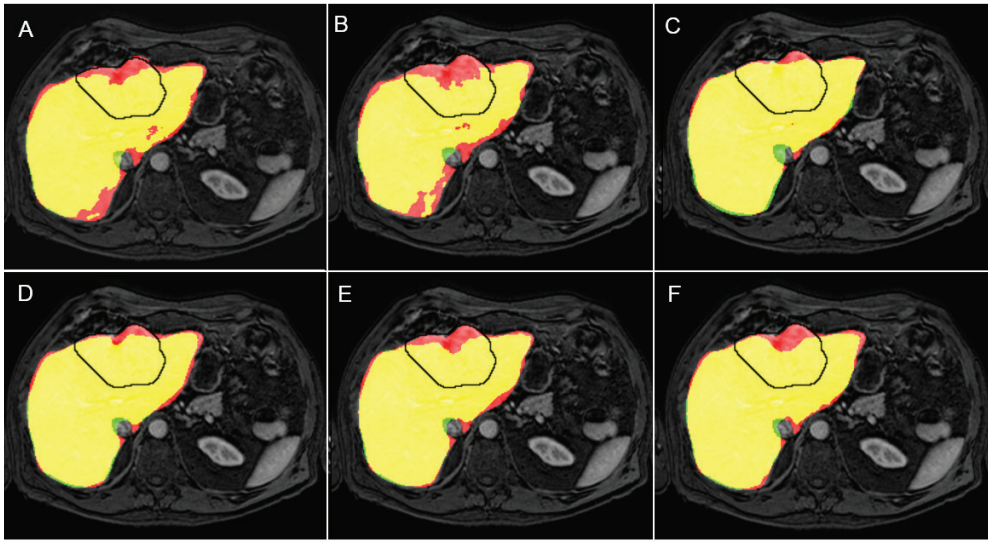


Figure 5 Segmentation results of a liver with a lesion, same patient as depicted in Figure 4 (left). The first row represents the dilated FCN and the second row the modified U-net. From left to right the columns represent configurations I, II, and III. The liver lesion is contoured in black. Yellow is true positive, red is false negative, and green is false positive.

Discussion and conclusions

Combining the different phases of the DCE-MR series in different channels results in significantly better segmentations than combining the feature maps of the different phases at the end of the network. It seems that the network learns the relationship between the phases of the DCE-MR series better when the intensity values still represent the MR intensity, rather than when several convolutional kernels are already applied. Especially the small U-net seems to fail to combine the feature maps at the second last convolutional layer in a proper way. Furthermore, including more information in the input images provides significantly better segmentation results than without the extra information for the dilated FCN.

For both networks, the number of learned parameters is in a similar order of magnitude for the three input configurations, which means the risk of overfitting is similar for all approaches[10] and leads to a fairer comparison between the performances of three configurations. However, this meant that the weights of the convolutional kernels needed to be shared in the networks with configuration II. This might not be the most optimal setup, because separately optimizing the kernel weights for each of the six phases of the DCE-MR series might result in better segmentation results. This may explain why configuration II performs less in some cases than configuration I, although more information is provided.

The results did not show a significant difference between the two network architectures when using the six phases as channels input image (configuration II). This suggest that the input configuration is more important than the network architecture. Visual inspection showed that, in general, the dilated FCN tends towards oversegmentation, while the modified U-net tends towards undersegmentation of the liver.

The extreme outliers of the DSC in the dilated FCN with configuration II are due to severe undersegmentation. The extreme outliers of the modified HD in the dilated FCN with configuration III are due to oversegmentation into either the spleen or the heart.

The DSC and HD of the segmentation results differ significantly from the DSC and HD of the inter-observer agreement and intra-observer agreement. Nonetheless, an expert radiologist deemed the segmentations of input configuration III for both networks satisfactory as a rough outline of the liver for further image analysis. Optimizing both networks could boost the performance of the networks. In this feasibility study, both networks provide a rough liver segmentation, which included

most of the lesions, except for lesions near the edges of the liver, as can be seen in Figure 5.

In conclusion, the configuration of the input images has an important effect on the outcomes of convolutional neural networks. Using the six phases of a DCE-MR series as channels of an input image performs better than using one phase image as input or the six phase images separately as input, regardless of the architecture of the network.

References

1. Lopez-Mir, F., Naranjo, V., Angulo, J., Alcañiz, M. & Luna, L. Liver segmentation in MRI: A fully automatic method based on stochastic partitions. *Comput. Methods Programs Biomed.* **114**, 11–28 (2014).

2. Huynh, H. T., Le-Trong, N., Bao, P. T., Oto, A. & Suzuki, K. Fully automated MR liver volumetry using watershed segmentation coupled with active contouring. *Int. J. Comput. Assist. Radiol. Surg.* **12**, 235–243 (2017).

3. Chartrand, G. *et al.* Liver Segmentation on CT and MR Using Laplacian Mesh Optimization. *IEEE Trans. Biomed. Eng.* **64**, 2110–2121 (2017).

4. Masoumi, H., Behrad, A., Pourmina, M. A. & Roosta, A. Automatic liver segmentation in MRI images using an iterative watershed algorithm and artificial neural network. *Biomed. Signal Process. Control* **7**, 429–437 (2012).

5. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9351**, 234–241 (2015).

6. Yu, F. & Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. in *ICLR* (2016). doi:10.16373/j.cnki.ahr.150049

7. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

8. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. in *Fourth International Conference on 3D Vision (3DV), 2016* 1–11 (2016). doi:10.1109/3DV.2016.79

9. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010* **9**, 249–256 (2010).

10. Jain, A. K., Duin, R. P. . & Mao, J. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 4–37 (2000).

4

Liver segmentation and metastases detection in MR images using convolutional neural networks

Based on:
M.J.A. Jansen, H.J. Kuijf, M. Niekel, W.B. Veldhuis, F.J. Wessels, M.A. Viergever, and J.P.W. Pluim, "Liver segmentation and metastases detection in MR images using convolutional neural networks", *Journal of Medical Imaging* 2019, 6(4): 044003

Abstract

Primary tumors have a high likelihood of developing metastases in the liver and early detection of these metastases is crucial for patient outcome. We propose a method based on convolutional neural networks (CNN) to detect liver metastases. First, the liver was automatically segmented using the six phases of abdominal dynamic contrast enhanced (DCE) MR images. Next, DCE-MR and diffusion weighted (DW) MR images are used for metastases detection within the liver mask. The liver segmentations have a median Dice similarity coefficient of 0.95 compared with manual annotations. The metastases detection method has a sensitivity of 99.8% with a median of 2 false positives per image. The combination of the two MR sequences in a dual pathway network is proven valuable for the detection of liver metastases. In conclusion, a high quality liver segmentation can be obtained in which we can successfully detect liver metastases.

Introduction

Primary tumors, such as neuroendocrine and colorectal tumors, have a high likelihood of developing metastases in the liver.[1,2] Early detection of (new) liver metastases is crucial since it improves patient outcome.[3-5] To follow disease progress, radiologists check for tumor growth and new (liver) metastases in computed tomography (CT) or magnetic resonance (MR) images.

While CT has long been the modality of choice in detecting and monitoring liver tumors, MRI has gained interest thanks to a better lesion-to-liver contrast and because it does not use ionizing radiation.[6,7] Dynamic contrast enhanced (DCE) MR images have a high sensitivity and specificity for visual detection of liver metastases. Moreover, the combination of DCE-MR and diffusion weighted (DW) MR images turned out to be even more effective in the visual detection of liver metastases and visual censoring of mimics.[8,9,10]

The automatic detection and characterization of liver metastases remains a challenging task, given the heterogeneous appearance of liver metastases on MR images.[11,12] Radiologists have a liver lesion detection rate between 87-95%, when using both DCE-MR and DW-MR images.[7,8,9] Automatic liver metastases detection could aid radiologists in finding metastases more efficiently and effectively.

The liver metastases detection method we propose is a two-step method: a liver segmentation step followed by a lesion detection step within the segmented liver.

Most liver segmentation methods published have been developed for CT, with best results by convolutional neural network (CNN) based methods.[13,14] Fewer methods have been developed for segmentation of the liver in MR images. Those known are primarily based on watershed[15,16], active contouring[17], atlases[18], or shape models[19].

For automatic detection of liver lesions several methods have been proposed, all based on CT images.[14, 20-23] MR data was employed by a few methods to detect hepatocellular carcinomas (either DCE-MRI[24] or DW-MRI[25]), but not for metastases.

Our aim is to develop and evaluate a two-step liver metastases detection method for MR images, based on fully convolutional neural networks (FCN). In the first step, a liver segmentation method utilizing the dynamic nature of the DCE-MR images is presented, based on previous work[26]. This is followed by a method for the detection of metastases within the liver region using both DCE-MR and DW-MR images as input.

Data

The study comprises MR data of 121 patients with a clinical focus on the liver from the University Medical Center Utrecht, the Netherlands, acquired between February 2015 and February 2018. The UMCU Medical Ethical Committee has reviewed this study and informed consent was waived due to its retrospective nature.

All patients underwent a clinical MR examination, including DCE-MR series and DW-MRI. The DCE-MR series was acquired in six breath holds with one to five 3D images per breath hold, with the following parameters: TE: 2.143 ms; TR: 4.524 ms; flip angle: 10 degrees. After acquiring the first image, gadobutrol (0.1 ml/kg Gadovist of 1.0 mmol/ml at 1 ml/s) was administered at once, followed by 25 ml saline solution at 1 ml/s. In total, 16 3D images per patient were acquired with 100 slices and matrix sizes of 256 ×256. Voxel size was 1.543 mm ×1.543 mm ×2 mm.

The DW-MR images were acquired with three b-values: 10, 150, and 1000 s/mm2, using a protocol with the following parameters: TE: 70 ms; TR: 1.660 ms; flip angle: 90 degrees. Each b-value image was acquired with 42 slices and matrix sizes of 256 × 256. Voxel size was 1.758 mm × 1.758 mm × 5 mm.

Pre-processing

DCE-MR

All DCE-MR data sets were corrected for motion using a groupwise registration.[27] The groupwise registration method registers all images simultaneously to a common space by minimizing a cost function based on principle component analysis and applying a B-spline transformation. The registration was applied on four resolutions with 500 iterations. After registration, a zero-mean-unit-variance rescaling was applied to all intensity values between the 0^{th} and 99.8^{th} percentile of the intensity histogram of DCE-MR series. The 99.8^{th} percentile intensity was assumed to correspond to the contrast agent peak in the aorta.

The images were combined per phase by averaging the fourth dimension. The series started with the pre-contrast image, followed by the other phases: the early arterial phase, the late arterial phase, the hepatic/portal-venous phase, the late portal-venous/equilibrium phase and the late equilibrium phase. The six phases were used as input images for the FCN.

DW-MR

The DW-MR images were nonlinearly registered to the DCE-MR space using elastix, a toolbox for linear and non-linear registration of medical images.[28] The fixed image was the mean DCE-MR, obtained by averaging over the six phases. First a rigid transformation was applied on two resolutions with 2000 iterations each, followed by a b-spline transformation on one resolution with 1000 iterations and a grid spacing of 60x60x40 mm. Normalized mutual information was used as metric.† A mask was used to focus the registration on the liver. The mask was obtained from the automatic liver segmentation, described in this chapter, morphologically dilated with a 10×10×10 structuring element. A substantial dilation was chosen to make sure the boundary of the liver is included.

Ten out of the 121 MRI examinations were excluded from the study because of a failed registration of the DW-MRI on the DCE-MR series. The exclusion was based on visual evaluation.

The intensities of the registered DW-MR data set were also normalized with a zero-mean-unit-variance rescaling between the 0^{th} and 99.8^{th} percentile of the intensity

† The parameter files used, are available at elastix.bigr.nl/wiki/index.php/Par0057

histogram of the DW-MRI.

Annotations

The liver was annotated in 55 DCE-MR series, of which only 16 series contained metastases that were segmented. Additionally, we included another 56 DCE-MR series in which the metastases were segmented, resulting in a total of 72 DCE-MR series with annotated metastases.

Liver segmentation

Fifty-five dynamic contrast enhanced MR series were used for liver segmentation. The data sets were randomly divided in thirty-three data sets for training, three for validation, and nineteen data sets for testing. The validation set was used for tuning the hyperparameters and the evaluation of the CNN model during method development. The test set was used to evaluate the final CNN model in an independent manner.

The liver was manually contoured by two observers, see Fig. 3 for segmentation examples. The first observer annotated the training, validation, and test sets. The annotation of the test set is indicated as O1.1. The first observer repeated the annotations of the test set at least one week later (O1.2). The second observer annotated the test set once (O2). This was done to estimate the inter- and intra-observer agreement. A radiologist with more than ten years of experience in liver MR analysis verified all manual annotations and provided corrections where needed. Liver lesions were included in the annotation and this network was therefore trained to recognize liver lesions as liver tissue. The first set of annotations of the first observer was used as reference in the experiments.

Liver metastases detection

Seventy-two MR data sets were used for the liver metastases detection. The data set included mainly colorectal metastases, neuroendocrine metastases, and some other metastasis types (i.e. other gastrointestinal metastases and breast metastases). The data sets were randomly divided in fifty-five data sets for training (n=50) and validation (n=5), and seventeen data sets for testing. The training and validation sets contained 334 metastases in total, with on average 6 metastases per liver (range: 1 - 31 metastases). The test set contained 86 metastases in total, with on average 5 metastases per liver (range: 1 - 32 metastases).

The metastases were manually annotated on the DCE-MR images by a radiologist in training and were verified by a radiologist with more than ten years of experience.

Methods

Liver segmentation

An FCN[29] with dilated convolutions was implemented, for which the six phases of the DCE-MR images were the channels of the input image. The dilated FCN consisted of 9 convolutional layers in total. The first 7 layers had a 3×3 convolution and 32 kernels. The dilation rates ranged from 1 to 16. The final two layers had a 1×1 convolution. ReLU activation and batch normalization were used in all the convolutional layers, except for the final layer which had a softmax activation. A dropout layer was applied between the 8th and the 9th layer with a dropout rate of 0.5. This network had a receptive field of 67×67 pixels. Figure 1 gives an overview of the network architecture. In our previous work on liver segmentation[26], we also considered the popular U-net architecture. However, this architecture had a slightly worse performance than the herein used FCN architecture.

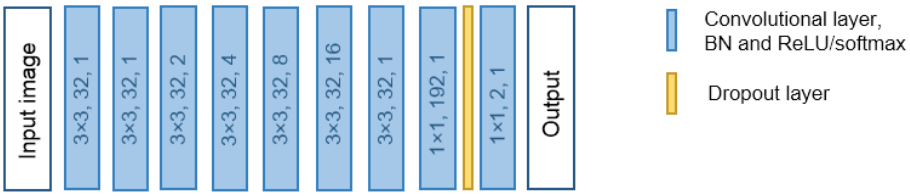


Figure 1 Network architecture for liver segmentation. The blue blocks represent the convolutional layers, batch normalization (BN) and ReLU or softmax activation. The size of the kernel of each convolutional layer is given in the block, followed by the number of kernels and the dilation rate of the kernel.

The loss was calculated by a similarity metric based on the Dice similarity coefficient: $\frac{2(X \cap Y)+s}{X^2+Y^2+s}$, where X is the predicted segmentation, Y is the ground truth mask, and s is a small number to prevent dividing by zero (set to 1e-5).[30] Glorot uniform[31] was used as initializer and Adam as optimizer with a learning rate of 0.001. The network was trained for 100,000 iterations, with six images per mini batch. The total number of 2D slices for training was 3300 slices. No data augmentation was applied.

The data was processed by the network per 2D slice consisting of 6 channels. For the evaluation of the liver segmentation, the probability output was post-processed to a binary image. The threshold of 0.5 was applied to the probability output of the network. After that 3D hole filling was performed, so that all holes caused by liver

lesions were filled. To remove small spurious segmentations, the largest connected component was selected as the final segmentation.

Liver metastases detection

A dual pathway FCN was implemented, for which the input images of one path were the six DCE-MR phase images and for the other path the three DW-MR images. The six (or three) 2D images were combined to one input image, with six (or three) channels. Each pathway had 13 convolutional layers with a 3×3 convolution and 64 kernels, split in five blocks with different dilation rates, ranging from 1 to 8. The feature maps at the end of each block of each pathway were concatenated in the third dimension, resulting in a feature map with 640 kernels, and were passed to two convolutional layers with a 1×1 convolution with 128 and 2 kernels, respectively. This resulted in a receptive field of 123×123 pixels. The individual pathways were inspired by the P-net architecture.[32] Figure 2 gives an overview of the network architecture.

In addition, the network was also trained and evaluated as a single pathway FCN, once with only the DCE-MR images as input images with six channels, and once with the DCE-MR and DW-MR images concatenated as input images with nine channels. In this manner the additional value of the DW-MR images and the addition of a second pathway in the detection method can be determined. The single pathway FCN was identical to the top pathway in Figure 2. In the concatenation layer, the feature maps at the end of each block were concatenated in the third dimension.

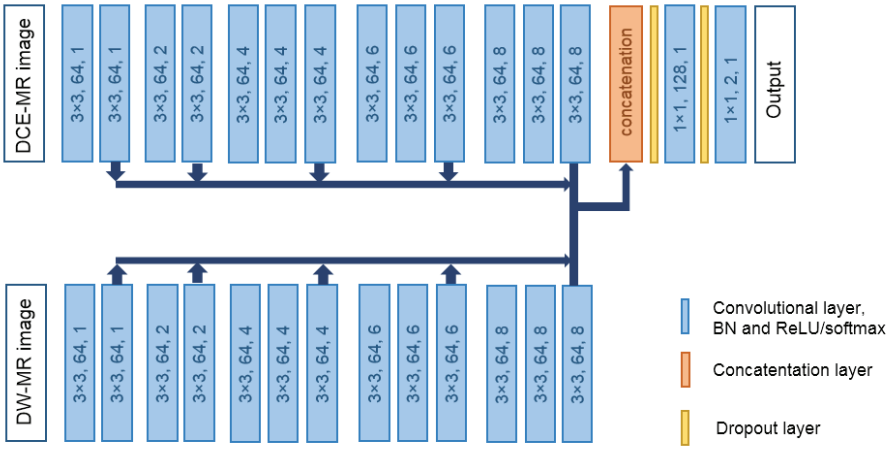


Figure 2 Network architecture for liver metastases detection. The blue blocks represent the convolutional layers, batch normalization (BN) and ReLU or softmax activation. The size of the kernel of each convolutional layer is given in the block, followed by the number of kernels and the dilation rate of the kernel.

Categorical cross-entropy was used as the loss function. ReLU activation and batch normalization were used in all the convolutional layers, except for the final layer which had a softmax activation. Two dropout layers were applied before and after the second-to-last layer. The dropout rate was set to 0.2. The classes were weighted based on class frequencies. He uniform[33] was used as initializer and Adam as optimizer with a learning rate of 0.0001. The network was trained for 10,000 iterations, with 4 images per mini batch. The slices to train on were limited to the slices containing lesions for a more balanced data set, resulting in a total number of 1619 2D slices for training and 180 slices for validation.

Twenty-five patches of 128×128 pixels were taken from each slice for data augmentation. The patches originated from the liver region and have overlapping areas. Data augmentation was applied by random rotation of the patches, with rotation angles of ±45 degrees.

The data of the test set was processed by the network per 2D image slice. The input image slice consisted either of six channels of the DCE-MR phase images, three channels of DWI images, or nine channels, depending on the network. The probability output was masked by the liver segmentation, which was dilated with a 5×5 structuring element. The dilation of the liver segmentation is a safety measure to ensure that small failures in the liver segmentation would not lead to undetected liver metastases.

For the evaluation of the detection, the masked probability output was post-processed to a binary image. All pixels with a probability output higher than the threshold of 0.5 were labelled as metastasis. Morphological closing with a structuring element of 3×3×3 was applied to fill any holes in the binary image. The morphological closing was followed by a morphological opening with a plus-shaped structuring element of 3×3, to remove any remaining noise in the detection results. The resulting binary image was divided into separate objects representing individual metastases, using voxel clustering with 26-neighbourhood connection.

Experiments

Liver segmentation

The performance of the liver segmentation was evaluated based on the Dice similarity coefficient (DSC), the relative volume difference (RVD), and the modified Hausdorff distance (HD) at the 95th percentile. The first annotations of the first observer of

the test set (O1.1) are used to evaluate the automatic segmentation, since these were made by the same observer in the same session as the annotations of the training and validation sets.

DSC: $\frac{2\,X \cap Y}{X+Y}$, where X is the automatic segmentation and Y annotation O1.1.

RVD: $\frac{X-Y}{Y}$*100%

HD at 95th percentile: $\max\left(h_{95}(X,Y), h_{95}(Y,X)\right)$, with $h_{95}(X,Y) = K_{x\in X}^{95^{th}}\, min_{y\in Y}\|y-x\|$. Here, X is the set of boundary points $\{x_1, .. x_N\}$ of the automatic segmentation result, and Y is the set of boundary points $\{y_1, .. y_N\}$ of theannotation O1.1. K^{95th} denotes the 95th ranked value in the set of distances between all boundary points in X and the closest boundary points in Y.[34]

These three metrics were computed on the predicted segmentation results. In addition, the three metrics were computed for the second annotations of the first observer (O1.2) and the annotations of the second observer (O2) to obtain the intra- and inter-observer agreement, respectively.

Liver metastases detection

A liver metastasis was considered detected, and thus a true positive object, when the manual annotation and the predicted segmentations had an overlap greater than 0.

For the evaluation of the liver metastases detection, the true positive rate (TPR) and the number of false positives per case (FPC) were reported. The TPR was calculated as the number of true positive objects divided by the total number of true lesion objects. The FPC was calculated as the number of objects not overlapping with any true metastasis object. In addition, the TPR and FPC were given for several thresholds in a Free-response Receiver Operating Characteristic (FROC) curve.

The same metrics were applied to the segmentation results of the single pathway networks.

An expert radiologist verified the results and determined the underlying physiology of selected false positive objects.

Results

Figure 3 shows some examples of the late arterial phase of the DCE-MRI, the DW-MRI with b-value 150 s/mm², the manually annotated liver and metastases, and the automatic segmentation of the liver with the detection of the liver metastases involving both DCE-MR and DW-MR images in the dual pathway FCN. It shows good liver and lesion segmentations, with a false positive object in the last row, which is verified to be a cyst.

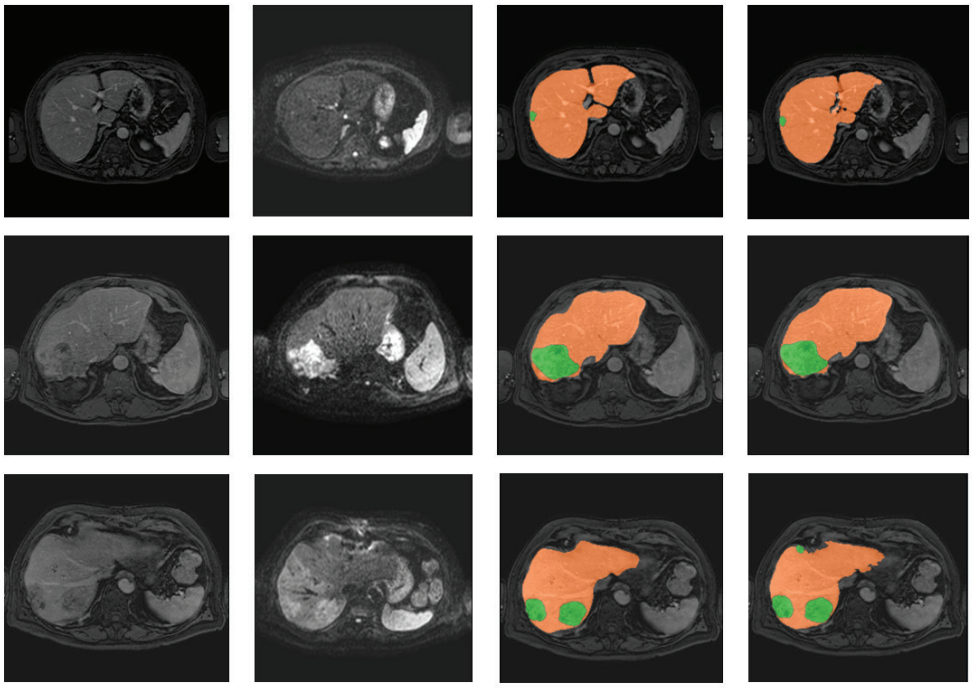


Figure 3 Three examples of liver and metastases segmentations. From left to right each column represents a late arterial phase DCE-MR image, the registered DW-MR image with b-value 150 s/mm², the manually annotated liver (orange) and metastases (green), and the automatically segmented liver (orange) and metastases (green). The bottom row shows a false positive object in the anterior side of the liver for the automatic detection, which is a cyst.

Liver segmentation

The median DSC is 0.95 for the automatic segmentation, 0.94 for the inter-observer agreement, and 0.96 for the intra-observer agreement. The median RVD is -1.6% for the automatic segmentation, -3.4% for the inter-observer agreement, and -1.5% for the intra-observer agreement. The median HD is 5.5 mm for the automatic segmentation, 5.6 mm for the inter-observer agreement, and 3.1 mm for the intra-observer agreement. The distributions of the DSC, RVD, and modified HD are given

in boxplots in Fig. 4.

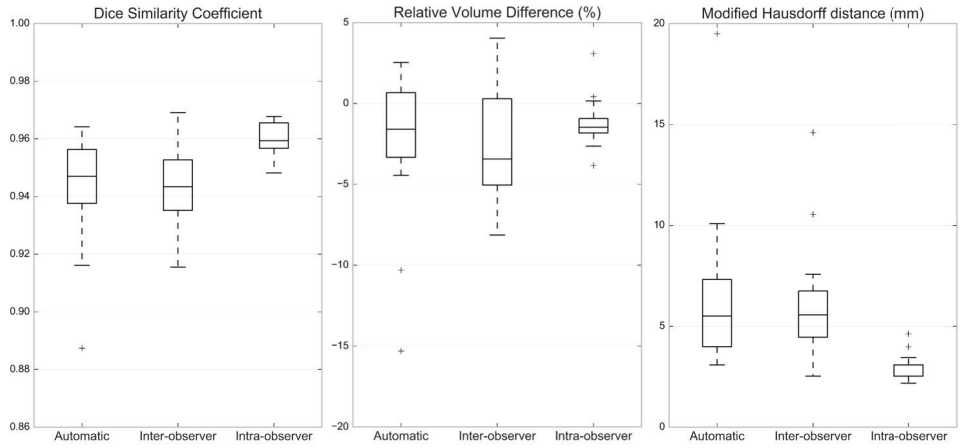


Figure 4 Boxplots of the Dice similarity coefficient, relative volume difference (%), and modified Hausdorff distance (mm).

In the boxplots, annotations O1.1 are used to evaluate the automatic liver segmentations. Some variance is present between the different annotations of the observers and therefore we also compared the automatic method to all manual annotations. The median [interquartile range (IQR)] values of the DSC, RVD, and modified HD for the automatic segmentation, the inter-observer agreements, and the intra-observer agreement using the three manual annotations are given in Table 1. The upper three rows give the same results as the boxplots and the lower three rows give additional results.

Example results of the automatic liver segmentation are shown in the last column of Figure 3 and in Figure 8.

Liver metastases detection

Figure 5 shows the TPR and the FPC for all subjects. The average TPR for the single pathway network using only DCE-MRI is 0.645 and for using both DCE-MRI and DW-MRI is 0.722. The average TPR is 0.998 for the dual pathway network using both DCE-MRI and DW-MRI as input images. The median FPC is, 5 for the single pathway network using only DCE-MRI, 6 when using both MR sequences, and 2 for the dual pathway network. The single pathway has 111 and 146 false positive objects in total, using only DCE-MRI and both MR sequences, respectively. While the dual pathway has 59 false positive objects in total.

Table 1 Median and interquartile range for the Dice similarity coefficient (DSC), relative volume difference (RVD), and modified Hausdorff distance (HD) are given. In each row of the first column, the latter observer is used as the reference. The upper three rows are the same results as given in Figure 4 and use O1.1 for evaluation. The lower three rows are additional results that either use O1.2 or O2 for evaluation. Note that O1.2 and O2 were created separately from the training data.

	DSC	RVD (%)	HD (mm)
Automatic – O1.1	0.95 [0.93 – 0.96]	-1.6 [-3.5 – 0.8]	5.5 [4.0 – 8.0]
Inter-observer O2 - O1.1	0.94 [0.93 – 0.95]	-3.4 [-5.1 – 0.6]	5.6 [4.3 – 6.8]
Intra-observer O1.2 - O1.1	0.96 [0.96 – 0.97]	-1.5 [-1.9 – -0.9]	3.1 [2.5 – 3.1]
Automatic – O1.2	0.95 [0.93 – 0.96]	-0.1 [-2.8 – 2.2]	4.9 [3.7 – 8.0]
Automatic – O2	0.95 [0.93 – 0.95]	0.5 [-3.8 – 5.5]	6.2 [4.9 – 9.6]
Inter-observer O2 - O1.2	0.95 [0.94 – 0.96]	-2.5 [-4.0 – 1.6]	4.4 [3.7 – 4.8]

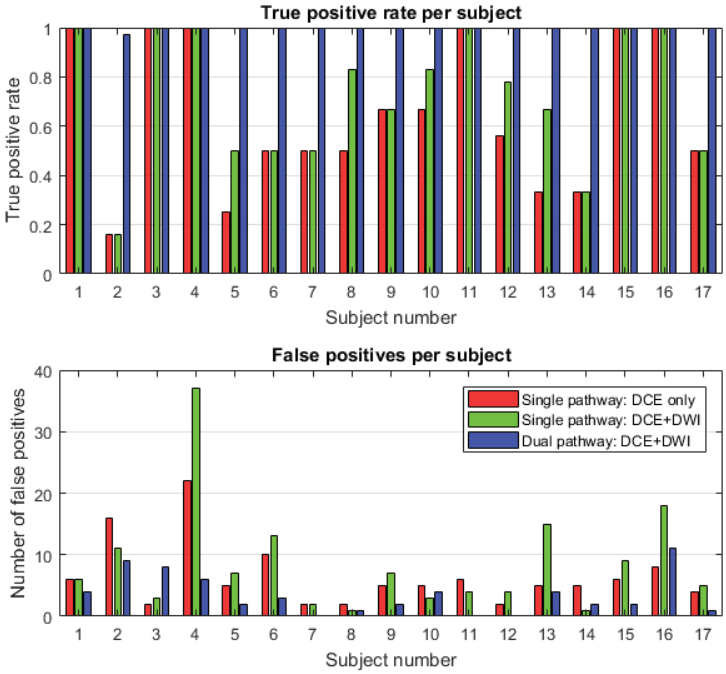


Figure 5 The true positive rate and the number of false positives per subject for the single pathway and the dual pathway networks.

Figure 6 shows the number of detected metastases relative to the total number of metastases per size category. Both the single pathway networks fail on the smaller metastases in particular.

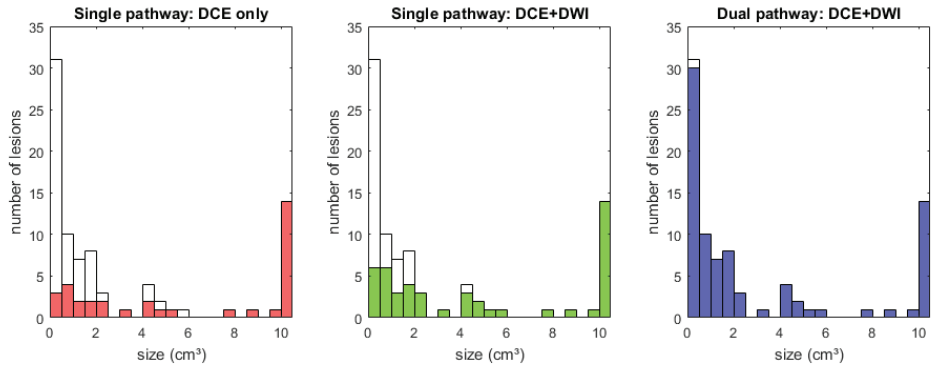


Figure 6 The number of lesions plotted against the size of the lesions. The colored part represents the detected lesions and the white part the missed lesions at a threshold of 0.50.

Figure 7 shows the FROC curve of the mean TPR versus the median FPC for different thresholds T applied to the output of both networks. The thresholds range from 0.90 to 0.00 with steps of 0.10. At the highest threshold (T=0.90) only 42% of the metastases was detected with a median FPC of one (in total, 31 false positive objects are present) using only the DCE-MR images, and 65% of the metastases was detected with a median FPC of one (in total, 25 false positive object are present) using both MR sequences in the single pathway. For the dual pathway, 96% of the metastases are detected, with a median FPC of one (in total, 19 false positive objects are present). . At the lowest threshold (T=0.0) 84% is detected with only DCE-MR images, 91% is detected with both MR sequences in the single pathway, and all the lesions are detected with the dual pathway network, but all at the cost of many false positives.

Figure 8 shows five examples of the lesion detection results including the liver segmentation result. The green pixels are true positive, blue pixels are false positive, and red pixels are false negative. Note that true positive objects are connected components that overlap with a manually annotated metastasis. A true positive object can nonetheless have some undersegmentation (red pixels, example indicated with arrow head) or oversegmentation (blue pixels, example indicated with arrow).

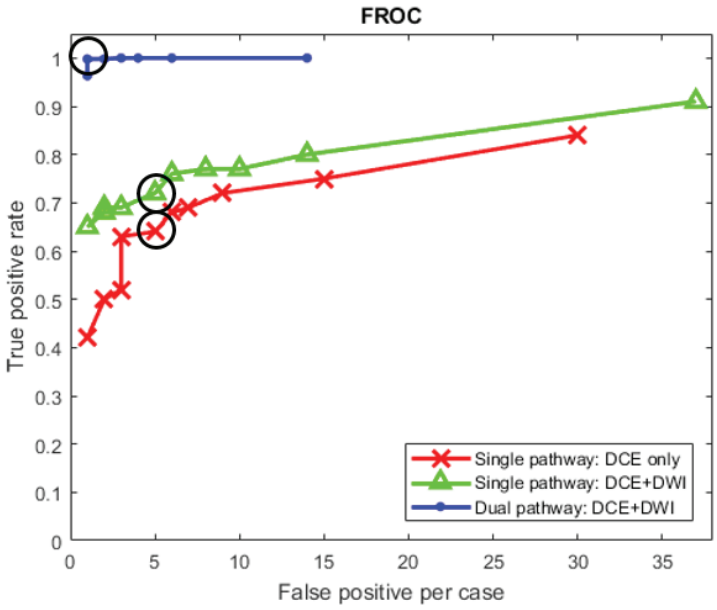


Figure 7 FROC curve of the mean true positive rate and the median number of false positives per case for threshold T ranging from 0.90 to 0.0 in steps of 0.10. The circles represent the TPR and FPC at threshold T=0.50.

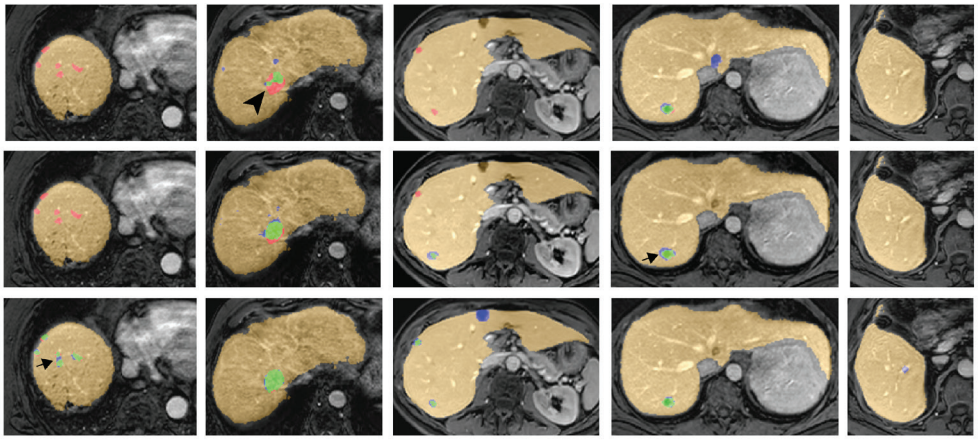


Figure 8 Five examples of the lesion detection results. The first row shows results of the single pathway using only DCE-MRI, the second row the results of the single pathway using DCE-MRI and DW-MRI, and the third row shows the results of the dual pathway network. Green pixels are true positive, red are false negative, blue are false positive pixels. The orange tissue represents the automatic liver segmentation. The arrows indicate oversegmentation and the arrow head indicates undersegmentation. From left to right the subjects 2, 4, 5, 10, and 17 are shown.

Discussion

Liver segmentation

Our proposed liver segmentation method is able to provide high quality liver segmentations using DCE-MR images. The method is able to deal with common difficulties in MR segmentation as well as the presence of lesions. The fact that the automatic liver segmentation results are not significantly different from the inter-observer results supports the claim of good liver segmentations.

The automatic method performs well on the test set of annotation O1.1. The results of the automatic method are similar to the inter-observer results, as can be seen in the top rows of Table 1. Observer 1 annotated the training set so it is not surprising that the automatic segmentations do well on the test set, which was annotated by the same observer, in the same annotation session. There is some intra-observer variance between O1.1 and O1.2 and the second annotations appear systematically smaller than the first annotations, since the RVD values involving O1.1 are mostly negative. Still, the automatic method also performs well when evaluated with annotations O1.2, as can be seen in the bottom rows of Table 1. Annotations O2 were used to calculate the inter-observer variance, and were additionally used to evaluate the automatic method with another observer to verify whether the method generalizes well. The results using O2 for evaluation are still good, although slightly worse than the automatic method evaluated with O1. This is as expected, because the first observer annotated the training and validation sets.

The automatic liver segmentation encountered few problems. However, an outlier in all three metrics was found in one case, where a large lesion (337 ml) occurred at the edge of the liver. This outlier has an RVD of -15%, as a result of undersegmentation of the liver in this lesion area. Even though the lesion is bigger than the receptive field of the neural network, the method accomplished to recognize half of the lesion tissue as part of the liver. However, large lesions at the edge of the liver are a potential pitfall for the proposed liver segmentation method.

It is hard to compare the liver segmentation results with other methods, since the data sets and MRI sequences used are not the same. A rough comparison with recently proposed automatic liver MR segmentation algorithms[17,18,25] can be made by considering the reported values for the DSC. For the mentioned studies, the DSC ranged from 0.87 to 0.91, while we report a DSC of 0.95.

Liver metastases detection

The detection of all liver metastases is of great importance for adequate treatment.[35] The detection method with only DCE-MR images has an average TPR of 0.645 and a median of 5 false positives per case. The detection method with the single pathway using both MR sequences is more effective, with an average TPR of 0.722 and a median of 6 false positives per case. The dual pathway detection method is able to detect almost all metastases at a threshold level of 0.50 with a median of 2 false positives per case. This shows the importance of adding DW-MR images to the lesion detection method and processing the two MR sequences in separate pathways. DW-MR images are highly sensitive for lesions and in combination with the DCE-MR images the method is able to detect metastases at a high detection rate and a low number of false positives. Fig. 8 shows some visual examples of the differences in detection of these three networks.

The usage of the two MR sequences requires the registration of the DW-MR images to the DCE-MR image space. Ten registrations failed, which led to the exclusion of these images from the study. This is likely caused by the free-breathing protocol used for the DW-MR images that may result in deformations that cannot be corrected by the registration method. A more dedicated registration approach or a breath-hold DW-MR acquisition may improve the robustness of the method.

One small metastasis is missed by the dual pathway network. The metastasis is only 5 pixels in size and was detected by the dual pathway with a probability higher than 0.50. However, it is deleted during the binary opening in the post-processing step, where it is seen as noise. This reduction of noisy false positives comes at the cost of missing very small metastases.

The missed metastases by the single pathway networks are mostly smaller than 2 cm^3, as can be seen in Fig. 6. These smaller metastases may differ in appearance from larger ones, because certain features, such as rim enhancement, cannot be expressed in only a few voxels of the DCE-MR images. The network with only DCE-MR input images is uncertain about these metastases, which leads to low outcome probabilities. However, the single pathway with both DCE-MR and DWI-MR is also less able to detect these smaller metastases. The network seems to fail to create adequate feature maps to detect small metastases, when it combines the two MR sequences in the first convolutional layer. The dual pathway network is allowed to create feature maps for the two MR sequences separately. The phases of the DCE-MR images are correlated over time, as are the three b-value images of the DW-MR. Both MR sequences

contain useful characteristics. In the dual pathway network, the feature maps are solely composed of the MR sequence presented to that pathway, representing those characteristics. And the feature maps are combined at the end. This might explain the difference the detection of the small metastases of the two networks. In addition, the dual pathway has two times more parameters as the single pathway. Fig. 8, first column, shows some examples of sub-centimeter sized metastases.

The dual pathway network detects almost all metastases, but also incorrectly marks other objects as metastases. Most of these false positives are caused by objects which are not or scarcely represented in the training set and have thus an appearance unknown to the network. These unknown appearances could be other lesions and liver conditions, such as cirrhosis or an inhomogeneous fat distribution in the liver. The network has mostly seen healthy liver parenchyma with metastases during training and has learned to distinguish these, but is uncertain about conditions and lesions not seen during training. Fig. 3 (last row) and Fig. 8 (third column) show examples of cysts marked as a metastasis by the detection method.

Inspection of the 19 false positive objects for a threshold level of 0.90 in the post-processing step reveals that 10 of the false positives are lesions: 9 cysts and 1 hemangioma. Three of the false positives are blood vessels and one false positive is because of an inhomogeneous fat distribution in the liver. Fig. 8 (last column) shows an example of a blood vessel marked as a metastasis. Three other false positives are located on the edge of the liver, which sometimes has a higher intensity than the rest of the liver (e.g. Fig. 3 second column). Furthermore, motion artifacts can be the cause of a false positive object, which is the case for two false positives.

The morphological operations of section 3.2 were carefully designed for the tasks described. However, post hoc analysis revealed that the liver mask dilation with a 5×5 structuring element did not seem to be necessary for this data set, since all liver metastases were included in the original liver mask. On the other hand, the morphological closing and opening of the lesion detection results did have an impact on the results of the dual pathway method. Without these morphological operations the median FPC would increase to 8 instead of 2. Nevertheless, the morphological operations removed one small lesion, from a total of 86, which was otherwise detected.

These CNN models are trained on images from a specific MR protocol from one clinic. Like other CNN models, there might be a drop in performance when the model is used on data from another scanner or another clinic[36]. To avoid this problem, the CNN should be fine-tuned on data similar to the test data, e.g. by using transfer

learning.

This work could also be expanded to 3D input images, adding more spatial information, which might improve the results. However, this would require more training data and computational power to train and test a network with more parameters.

Conclusions

An automatic liver segmentation with a similarity index comparable to that of the inter-observer agreement, is obtained from dynamic contrast enhanced MR images with a fully convolutional neural network. The method can accurately segment livers irrespective of the liver condition or the presence of lesions. Only in the event of lesions larger than the receptive field of the fully convolutional neural network, parts of the liver are missed.

The proposed dual pathway metastases detection method, based on dynamic contrast enhanced MR and diffusion weighted MR images, successfully detects 99.8% of the liver metastases at the cost of a median of two false positives per case. This will aid radiologists to locate metastases quickly.

References

1. Vogl, T. J. *et al.* Liver metastases of neuroendocrine carcinomas: Interventional treatment via transarterial embolization, chemoembolization and thermal ablation. *Eur. J. Radiol.* **72**, 517–528 (2009).

2. Van Cutsem, E. *et al.* Towards a pan-European consensus on the treatment of patients with colorectal liver metastases. *Eur. J. Cancer* **42**, 2212–2221 (2006).

3. Robinson, P. J. The early detection of liver metastases. *Cancer Imaging* **2**, 1–3 (2002).

4. Bakhtiary, Z. *et al.* Targeted superparamagnetic iron oxide nanoparticles for early detection of cancer: Possibilities and challenges. *Nanomedicine Nanotechnology, Biol. Med.* **12**, 287–307 (2016).

5. World Health Organization – International Agency for Cancer Research. *GLOBOCAN 2018, Cancer Incidence, Mortality and Prevalence Worldwide in 2018.* (2018).

6. Silva, A. C. *et al.* MR imaging of hypervascular liver masses: a review of current techniques. *Radiographics* **29**, 385–402 (2009).

7. Böttcher, J. *et al.* Detection and classification of different liver lesions : Comparison of Gd-EOB-DTPA-enhanced MRI versus multiphasic spiral CT in a clinical single centre investigation. *Eur. J. Radiol.* **82**, 1860–1869 (2013).

8. Vilgrain, V. *et al.* A meta-analysis of diffusion-weighted and gadoxetic acid-enhanced MR imaging for the detection of liver metastases. *Eur. Radiol.* **26**, 4595–4615 (2016).

9. Holzapfel, K. *et al.* Detection , classification , and characterization of focal liver lesions : Value of diffusion-weighted MR imaging , gadoxetic acid-enhanced MR imaging and the combination of both methods. *Abdom. Imaging* **37**, 74–82 (2012).

10. Kenis, C. *et al.* Diagnosis of liver metastases : Can diffusion-weighted imaging be used as a stand alone sequence? *Eur. J. Radiol.* **81**, 1016–1023 (2012).

11. Elsayes, K. M. *et al.* Focal hepatic lesions: Diagnostic value of enhancement pattern approach with contrast enhanced 3D gradient-echo MR imaging. *Radiographics* **25**, 1299–1320 (2005).

12. Schmid-Tannwald, C. *et al.* Diffusion-weighted MRI of metastatic liver lesions: is there a difference between hypervascular and hypovascular metastases ? *Acta radiol.* **55**, 515–523 (2014).

13. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

14. Christ, P. F. Liver Tumor Segmentation Challenge. (2017). Available at: lits-challenge.com.

15. Lopez-Mir, F., Naranjo, V., Angulo, J., Alcañiz, M. & Luna, L. Liver segmentation in MRI: A fully automatic method based on stochastic partitions. *Comput. Methods Programs Biomed.* **114**, 11–28 (2014).

16. Masoumi, H., Behrad, A., Pourmina, M. A. & Roosta, A. Automatic liver segmentation in MRI images using an iterative watershed algorithm and artificial neural network. *Biomed. Signal Process. Control* **7**, 429–437 (2012).

17. Huynh, H. T., Le-Trong, N., Bao, P. T., Oto, A. & Suzuki, K. Fully automated MR liver volumetry using watershed segmentation coupled with active contouring. *Int. J. Comput. Assist. Radiol. Surg.* **12**, 235–243 (2017).

18. Dura, E., Domingo, J., Göçeri, E. & Martí, L. A method for liver segmentation in perfusion MR images using probabilistic atlases and viscous reconstruction. *Pattern Anal. Appl.* **21**, 1083–1095 (2018).

19. Chartrand, G. *et al.* Liver Segmentation on CT and MR Using Laplacian Mesh Optimization. *IEEE Trans. Biomed. Eng.* **64**, 2110–2121 (2017).

20. Ben-Cohen, A. *et al.* Fully convolutional network and sparsity-based dictionary learning for liver lesion detection in CT examinations. *Neurocomputing* **275**, 1585–1594 (2018).

21. Christ, P. F. *et al.* Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. in *Lecture Notes in Computer Science MICCAI 2016* **9901**, 415–423 (2016).

22. Bilello, M. *et al.* Automatic detection and classification of hypodense hepatic lesions on contrast-enhanced venous-phase CT. *Med. Phys.* **31**, 2584–2593 (2004).

23. Vorontsov, E., Tang, A., Pal, C. & Kadoury, S. Liver lesion segmentation informed by joint liver segmentation. in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 1332–1335 (2018).

24. Pavan, A. L. M. *et al.* A Parallel Framework for HCC Detection in DCE-MRI Sequences with Wavelet-Based Description and SVM classification. in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* 14–21 (2018). doi:10.1145/3167132.3167167

25. Christ, P. F. *et al.* Automatic Liver and Tumor Segmentation of CT and MRI Volumes Using Cascaded Fully Convolutional Neural Networks. *arXiv Prepr. arXiv1702.05970* 1–20 (2017).

26. Jansen, M. J. A., Kuijf, H. J. & Pluim, J. P. W. Optimal input configuration of dynamic contrast enhanced MRI in convolutional neural networks for liver segmentation. in *Proc. SPIE 10949, Medical Imaging 2019: Image Processing* 1094941V (2019). doi:10.1117/12.2506770

27. Jansen, M. J. A. *et al.* Evaluation of motion correction for clinical dynamic contrast enhanced MRI of the liver. *Phys. Med. Biol.* **62**, 7556–7568 (2017).

28. Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. W. elastix: A toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* **29**, 196–205 (2010).

29. Yu, F. & Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. in *ICLR* (2016). doi:10.16373/j.cnki.ahr.150049

30. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. in *Fourth International Conference on 3D Vision (3DV), 2016* 1–11 (2016). doi:10.1109/3DV.2016.79

31. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010* **9**, 249–256 (2010).

32. Wang, G. *et al.* Interactive medical image segmentation using deep learning with image-specific fine-tuning. *IEEE Trans. Med. Imaging* **37**, 1562–1573 (2018).

33. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. in *Proceedings of the IEEE International Conference on Computer Vision* **2015**, 1026–1034 (2015).

34. Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**, 850–863 (1993).

35. Masi, G., Fornaro, L., Caparello, C. & Falcone, A. Liver metastases from colorectal cancer: how to best complement medical treatment with surgical approaches. *Futur. Oncol.* **7**, (2011).

36. Kuijf, H. J. *et al.* Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities; Results of the WMH Segmentation Challenge. *IEEE Trans. Med. Imaging* **99**, 1–1 (2019).

5

Patient-specific fine-tuning of CNNs for follow-up lesion quantification

Based on:
M.J.A. Jansen, H.J. Kuijf, A.K. Dhara, N.A. Weaver, G.J. Biessels, R. Strand, and J.P.W. Pluim, "Patient-specific fine-tuning of CNNs for follow-up lesion quantification" *Submitted*

Abstract

Convolutional neural network (CNN) methods have been proposed to quantify lesions in medical imaging. Commonly more than one imaging examination is available for a patient, but the serial information in these images often remains unused. CNN-based methods have the potential to extract valuable information from previously acquired imaging to better quantify current imaging of the same patient.

A pre-trained CNN can be updated with a patient's previously acquired imaging: patient-specific fine-tuning. In this work, we studied the improvement in performance of lesion quantification methods on MR images after fine-tuning compared to a base CNN. We applied the method to two different approaches: the detection of liver metastases and the segmentation of brain white matter hyperintensities (WMH).

The patient-specific fine-tuned CNN has a better performance than the base CNN. For the liver metastases, the median true positive rate increases from 0.67 to 0.85. For the WMH segmentation, the mean Dice similarity coefficient increases from 0.82 to 0.87. In this study we showed that patient-specific fine-tuning has potential to improve the lesion quantification performance of general CNNs by exploiting the patient's previously acquired imaging.

Introduction

Lesion quantification is an important step in medical image analysis. Over the years convolutional neural network (CNN) based lesion quantification methods have been proposed for medical images[1–5] to aid radiologists in the detection and segmentation of these lesions. For some diseases or treatments, there is a clinical need to follow-up on the patient. For these patients more than one imaging examination is available. The previously acquired scan, i.e. baseline scan, contains patient-specific information that could be exploited by the CNN-based method to better quantify the current scan of the same patient, i.e. the follow-up scan.

The conventional way to train a CNN is to iterate over a large data set representing the object to detect or segment. In this way, a method is obtained that should give reliable results in comparable data sets. However, CNN-based methods do not always provide satisfying performance for clinical practice. Differences in image quality and variations among patients are often not fully covered within the method and can cause insufficient results. As a result of training a CNN on a general population, the CNN is not fully adapted to the specific details of an individual patient.

Updating a CNN toward the features of a patient could improve the outcome of the CNN for that specific patient[6,7]. Previously acquired scans can be used for this purpose, as they are already analyzed at the time of examination of the follow-up scan. We therefore propose a method to enhance the performance of lesion quantification in a follow-up scan with a patient-specific fine-tuning approach. In addition, we gained insight on the necessary conditions to achieve these improvements.

Fine-tuning a pre-trained CNN has the benefit of obtaining good results using only a small data set during the fine-tuning step and has successfully been applied in previous studies. For example, medical image domain knowledge has been transferred to a CNN pre-trained on natural images by fine-tuning the CNN with medical images[8,9]. Pre-trained CNNs have been fine-tuned towards the features of one specific image, resulting in better segmentations of that image[10,11].

Patient-specific fine-tuning of a CNN is a method to update the CNN using a previous scan of a patient. This fine-tuned CNN has learned the specific features of abnormalities and the surrounding healthy tissue to improve the detection or segmentation in a follow-up scan. This approach is based on the assumption that the baseline and follow-up scan of a patient share the features of healthy tissue and abnormalities.

We present and evaluate the patient-specific fine-tuning approach on two applications; the detection of liver metastases on MRI and the segmentation of brain white matter hyperintensities (WMH) on MRI. Furthermore, different aspects of patient-specific fine-tuning are studied.

Materials and methods

The proposed lesion quantification framework is shown in Figure 1. First a CNN is trained using the training set, referred to as the base CNN model. Next, this base CNN is refined for each individual patient in the patient-specific fine-tuning step. A previous MRI scan of a patient, referred to as baseline scan, is used to fine-tune the network to the specific features of that patient and its lesions. During the testing step this patient-specific CNN model is used to detect or segment lesions in a follow-up MRI scan of the same patient. Two MR sequences are applied to train and test the CNN model.

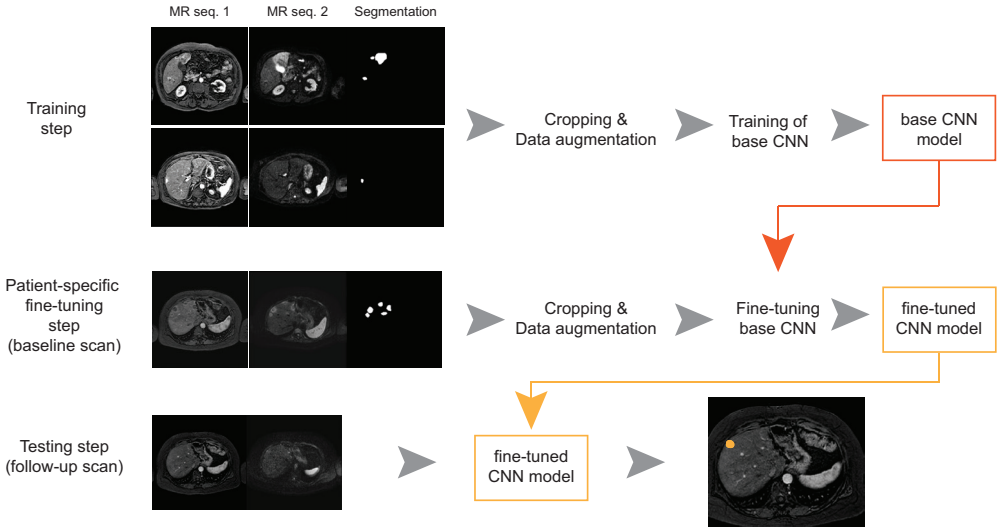


Figure 1 Proposed lesion quantification framework, shown with the liver MRI. First a base CNN is trained with a training set consisting of multiple patients. Next, the base CNN is refined in the patient-specific fine-tuning step using a previous MRI exam of a patient (the baseline scan). The fine-tuned CNN is used to detect or segment lesions in a follow-up MRI scan of the same patient. The images are cropped to focus of the organ of interest. The cropped image size is 128×128.

Data

The patient-specific fine-tuning approach is demonstrated on two different data sets: abdominal MRI for liver metastases detection and brain MRI for WMH segmentation.

Liver metastases detection

In this study, abdominal MRI of 47 patients with liver metastases from the University Medical Center Utrecht, the Netherlands, were included. Sixteen of them had at least two consecutive MRI examinations within a three months to a year. All patients underwent a clinical MRI examination, including dynamic contrast enhanced (DCE) MR series and diffusion weighted (DW) MRI, acquired on a 1.5T scanner (Philips, Best, the Netherlands). These two MR sequences are used to train the CNN. The UMCU Medical Ethical Committee has reviewed and approved this study and informed consent was waived due to its retrospective nature.

The DCE-MR series was acquired in six breath holds with one to five 3D images per breath hold, with in total 16 3D images. The DW-MR images were acquired with three b-values: 10, 150, and 1000 s/mm^2. The DCE-MR images were corrected for motion using a principle component analysis (PCA) based groupwise registration[12].

Intensity normalization was applied to both MR sequences and the DW-MR images were registered to the DCE-MR images by a rigid transformation, followed by a b-spline transformation‡. The resulting images have 100 slices and matrix sizes of 256 × 256. Voxel size is 1.543 mm × 1.543 mm × 2 mm.

The liver metastases were manually segmented on the DCE-MRI by a radiologist in training and verified by a radiologist with more than ten years of experience. The data set included mainly colorectal metastases, neuroendocrine metastases, and some other metastasis types (i.e. other gastrointestinal metastases and breast metastases). On average 30% of the liver slices contained liver metastases. Liver masks were automatically obtained using our previously developed segmentation method[13].

MRI data from 31 patients of which only one MRI examination was available were used to train the base CNN. The slices for training were limited to the slices containing metastases for a more balanced data set, resulting in a total number of 798 2D slices for training. The remaining 16 patients, with available baseline and follow-up scans, were used for testing. The scans had an average of 6 metastases per patient, ranging from 1 to 31 metastases.

WMH segmentation

The brain MRI data of 80 memory clinic patients with WMH were included in the study. Twenty of the patients were from the Dutch Parelsnoer Institute - neurodegenerative diseases study[14], from the UMC Utrecht, the Netherlands, and the remaining 60 were from the WMH challenge training set[15]. The 20 patients from UMC Utrecht had two MRI examinations, with two years between the two examinations.

All MR exams were acquired with a similar protocol, with a T2-weighted-Fluid-Attenuated Inversion Recovery (FLAIR) MRI and a T1-weighted MRI, acquired on a 3T scanner[16]. These two MR sequences are used to train the CNN. Both MR sequences were bias field corrected and the T1-weighted MRI were registered to the FLAIR images; more details can be found in Kuijf et al.[15]. All images were resized to a matrix size of 240 × 240 and 48 slices. Voxel size is 0.958 mm × 0.958 mm × 3.00 mm.

The WMH were manually segmented on the FLAIR images by an experienced researcher, in accordance with the STRIVE criteria[17]. WMH segmentation was performed with in-house developed software on MeVisLab (MeVis Medical Solutions AG, Bremen, Germany). On average 63% of the brain slices contained WMH. Brain

‡ The parameter files used, are available at elastix.bigr.nl/wiki/index.php/Par0057

masks were obtained using the SPM software[18].

The MRI data of 60 patients with a single MRI examination from the WMH challenge were used for training. The slices for training were limited to the slices containing WMH lesions for a more balanced data set, resulting in a total of 1383 2D slices for training the base CNN. The remaining 20 patients, with a baseline and follow-up scan, were used for testing. These scans had an average of 65 WMH per patient, ranging from 21 to 117 WMH.

Base CNN model

The overall architecture of the CNN was inspired by our earlier work on liver metastases detection[13]. This fully convolutional architecture includes elements from the P-net architecture, which has proven to be efficient for updating[10]. As most MRI exams usually consist of multiple MR sequences, we modified the original P-net to include a dual pathway that can process two MR sequences, each in a separate pathway to extract specific feature maps for those sequences. The input image for each pathway was one MR sequence. If the MR sequence had multiple instances, such as the phases of the abdominal DCE-MRI, the 2D images were combined into one input image with the instances as channels. Variations on the network architecture have been tested in our earlier work[13]. See Figure 2 for an overview of the fully convolutional network architecture.

Each pathway had 13 convolutional layers (2 convolutional layers and 11 dilated convolutional layers[19] with varying dilation rates), each having 3×3 convolution and 64 kernels, split in five blocks. The feature maps at the end of each block of each pathway were concatenated in the third dimension, resulting in a feature map with 640 kernels, and were passed to two convolutional layers with a 1×1 convolution with 128 and 2 kernels, respectively. This resulted in a receptive field of 123×123 pixels.

During training, categorical cross-entropy was used as the loss function. This loss function was calculated on a pixel level resulting in detection by segmentation. ReLU activation and batch normalization were used in all the convolutional layers, except for the final layer, which had a softmax activation. Two dropout layers were applied before and after the second-last layer. The dropout rate was set to 0.2. The classes were weighted based on class frequencies. The class weights were set to 1 for the background class and to 5 for the lesion class. He uniform was used as initializer and Adam as optimizer with a learning rate of 0.0001. The network was trained for 10,000 iterations, with 4 images per mini batch.

Twenty-five patches of 128×128 pixels were taken from each slice for data augmentation. The patches originated from the organ of interest and have overlapping areas. Online data augmentation was applied by random rotation of the patches, with rotation angles of ±45 degrees.

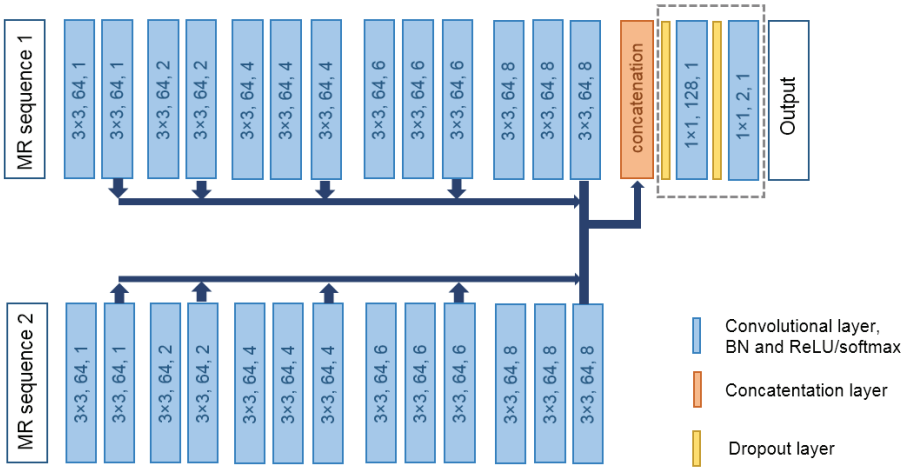


Figure 2 Fully convolutional network architecture for lesion quantification. The blue blocks represent the convolutional layers, batch normalization (BN) and ReLU or softmax activation. The size of the kernel of each convolutional layer is given in the block, followed by the number of kernels and the dilation rate of the kernel. The dashed line indicates the trainable layers during the patient-specific fine-tuning step.

Patient-specific fine-tuning

In the patient-specific fine-tuning step the last two layers of the base CNN were further trained with the baseline scan of a patient, to improve the lesion quantification results in the follow-up scan of the same patient. Only the last two layers were updated, as these combine the feature maps of the previous blocks of the two pathways and to ensure a fast fine-tuning step.

The CNN was refined by continuing training of the base CNN with the weights of all layers frozen, i.e. untrainable, except for those of the last two layers. The two trainable layers are indicated with the dashed line box in Figure 2.

The same loss function, optimizer, and learning rate were used as the ones during the training of the base CNN. The CNN was refined with 4 images per mini batch and the number of iterations was optimized during the experiments, see the Experiments section. The images used for fine-tuning were MRI slices of the patient's baseline scan.

The baseline scan was annotated earlier by an expert and only slices with at least one lesion present were included in the fine-tuning step. Five patches of 128×128 pixels were taken within the organ of interest region from each slice, four patches from the corners of the region and one center patch, removing air and tissues not of interest from the input images. The patches were randomly rotated, with rotation angles of ±45 degrees. Rotation of the patches included the variance in the positioning of the body during the MRI examination.

Lesion quantification on follow-up scan

The follow-up data was processed by the individual patient-specific CNN, i.e. the base CNN fine-tuned with the baseline scan of that patient. The probability output of the network was masked by the liver or brain mask.

For the evaluation, the masked probability output was post-processed to a binary image. A threshold of 0.5 was applied to the softmax output of the network and morphological closing with a structuring element of 3×3×3 was applied to fill holes. For the liver data set, the morphological closing was followed by a morphological opening with a plus-shaped structuring element of 3×3, to remove noise pixels. This was not applied to the brain data set as WMHFF of 1 pixel occur frequently. For the liver data set, the resulting binary image was divided into separate objects representing individual lesions, using voxel clustering with 26-neighbourhood connection.

Experiments

The setup for the base CNN model and the process of fine-tuning the CNN, as explained above, are similar for both the liver metastases detection and the WMH segmentation. The same experiments are conducted, only the evaluation is specified for the task: detection or segmentation.

Liver metastases detection

The detection of (new) liver metastases is important to monitor disease progression and for treatment planning. Treatment selection is based on the detection findings, i.e. the location and the number of metastases. Progress is monitored by follow-up scans.

The detection efficacy is evaluated using the true positive rate (TPR), the number of false positives per case (FPC), and the F1 score. As the values of the TPR, FPC, and

F1 score do not have a normal distribution, the median and interquartile range (IQR) are reported and the Wilcoxon signed rank test will be used to test for significant differences.

The TPR is calculated as the number of true positive objects divided by the total number of true lesion objects. A lesion is considered detected, and thus a true positive object, when the manual annotation and the predicted segmentations have an overlap greater than 0. The FPC is calculated as the number of detected objects not overlapping with any true lesion object. The F1 score is calculated as $\frac{2*recall*precision}{recall+precision}$, where *recall* is the TPR as defined above and *precision* is the number of true positive objects divided by the total number of detected objects (both true positive and false positive objects).

WMH segmentation

WMH are a common radiological finding in the elderly population, and are generally considered to reflect cerebral small vessel disease.[20] The presence and extent of WMH are associated with cognitive decline and dementia.[17,20,21] In particular, progression of WMH over time has been linked to cognitive decline and risk of dementia.[22,23] Quantification of WMH volume changes over time could contribute to the monitoring of disease progression, and provide clinicians with relevant information for informed decisions.

WMH segmentation on brain MRI is evaluated per MRI exam using the overall Dice Similarity Coefficient (DSC) and the absolute volume difference (AVD). The Dice score is calculated as $\frac{2*X \cap Y}{X+Y}$, where X is the automatic segmentation and Y the manual annotation. The AVD is calculated as $\frac{abs(V(X)-V(Y))}{V(Y)} * 100\%$, where $V(X)$ is the automatic segmentation volume and $V(Y)$ the manual annotation volume. The mean and standard deviation are reported and the paired Student's t-test is used to test for significant differences.

The following experiments study different elements of patient-specific fine-tuning; the number of iterations, the number of slices presented during fine-tuning, and a weighting scheme.

Number of iterations

The duration of the fine-tuning should be adapted to the similarity between the baseline and the follow-up scan, as the baseline scan can have a different appearance to the follow-up scan. For example, liver metastases show changes in shape and

size after treatment, and WMH are known to progress from punctuate to confluent lesions over time. If the CNN is fine-tuned for too few iterations the CNN will not be patient specific enough. If the number of iterations is too high, the risk of overfitting towards the baseline scan increases. The number of iterations is therefore studied and the optimal number will be used in further experiments. The number of iterations explored ranges from 50 to 1000.

Number of slices

The possibility of fine-tuning with only one or two slices is studied. Annotating or correcting the annotation of the baseline scan on one or two slices would be a smaller effort than annotating a full image if a good annotation does not yet exist. In these experiments, the number of slices that are offered to the CNN are one, two, or all slices with lesions or the full organ of interest present.

When not all slices are included, selection is based on the softmax probability outcome of the base CNN on the baseline scan. The slices with an average softmax probability closest to 0.5 are selected, assuming that these slices contain the most valuable information for updating the CNN. Two options were explored for the slice selection procedure. The first is including all slices for slice selection and the second option is only including slices with lesions present for slice selection.

Weighting

Weighting false negatives, false positives, and true positives, obtained after performing the lesion quantification with the base CNN on the baseline scan, will redefine what we want the CNN to learn during fine-tuning. The weights are set according the task to be performed.

For the detection task, the objective is to detect the metastases on liver MRI. The weights of the missed metastases are set to the highest value (five). The weights of the true positive pixels in detected metastases are set to two, with the weights of false negatives and the false positive pixels connected to the detected metastases set to zero. The background and false positive objects weights are set to one. In this manner the fine-tuned CNN will be greedier in labelling pixels as liver metastasis. Over- or undersegmentation of detected metastases is not considered incorrect labelling and false positive objects are considered less problematic than missing a metastasis.

For the segmentation task, the objective is to segment the WMH on brain MRI and get a good lesion volume estimation. The weights of the false positive and false

negative pixels are set to the highest value (five), as the main goal is to reduce the incorrectly labelled pixels. The weights of the true positive pixels are set to two, and the true negative pixels are set to one, to handle class imbalance.

The range of values of the weights are based on the class weights during training of the base CNN, which were five and one for lesions and background, respectively.

Uncertainty of CNN

The uncertainty of the CNN in detecting or segmenting the lesions is assessed by the standard deviation (SD) in probability outcome of the softmax layer. We implemented Monte Carlo dropout during test time[24,25], repeating the test phase 25 times. The SD of the probabilities over the 25 repetitions is calculated for every voxel. For the detection task the mean SD of the detected metastases is calculated as uncertainty metric, to study the change in network uncertainty for detecting metastases. For the segmentation task the maximum SD of all voxels in an image is taken as uncertainty metric, to study the change in network uncertainty for the entire image.

Results

Liver metastases detection

Number of iterations

In Table 1 the TPR, FPC, and the F1 score are reported for different numbers of iterations during the fine-tuning step. For a duration of only 50 iterations the TPR is similar to the TPR of the base CNN, but for 100 iterations and higher the TPR improves. However, for more than 50 iterations the FPC increases slightly. No significant differences (Wilcoxon signed rank test with $p<0.01$) are found between the base CNN and the fine-tuned CNN for different number of iterations. For the remaining experiments, the CNN is fine-tuned for 100 iterations.

Number of slices

The median [IQR] of the TPR, FPC, and the F1 score for the inclusion of different (numbers of) slices are presented in Table 2. The slice selection considered either all liver slices or all slices containing metastases. The slices with an average softmax probability closest to 0.5 were selected.

Including all the slices with liver metastases for fine-tuning the CNN gives the best results for the liver metastases detection method. Fine-tuning the CNN with only one or two selected slices of the liver does not improve the results.

Table 1 Median [interquartile range] of the true positive rate (TPR), the false positive count (FPC) and the F1 score of the liver metastases detection, for a varying number of iterations of learning for the CNN for fine-tuning. The best results are printed in bold.

	TPR	FPC	F1 score
Base CNN	0.67 [0.32-1.00]	3 [0-7]	0.49 [0.21-0.67]
50 iterations	0.67 [0.20-1.00]	**1 [0-3]**	0.40 [0.25-0.80]
100 iterations	0.85 [0.44-1.00]	2 [0-3]	**0.57 [0.40-0.67]**
500 iterations	**0.92 [0.33-1.00]**	2 [0-4]	0.50 [0.25-0.71]
1000 iterations	0.85 [0.50-1.00]	1 [0-4]	0.50 [0.40-0.75]

Table 2 Median [interquartile range] of the true positive rate (TPR), the false positive count (FPC) and the F1 score for a ranging number of slices presented to the CNN for fine-tuning. The best results are printed in bold. No significant differences were found between the Base CNN and all options.

	TPR	FPC	F1 score
Base CNN	0.67 [0.32-1.00]	3 [0-7]	0.49 [0.21-0.67]
1 slice liver	0.50 [0.11-1.00]	4 [1-8]	0.33 [0.12-0.64]
2 slices liver	0.42 [0.11-1.00]	1 [0-5]	0.33 [0.15-0.67]
1 slice metastases	0.67 [0.08-1.00]	4 [2-6]	0.31 [0.00-0.50]
2 slices metastases	0.67 [0.08-1.00]	3 [1-5]	0.31 [0.00-0.44]
All metastases slices	**0.85 [0.44-1.00]**	2 [0-3]	**0.57 [0.40-0.67]**
All liver slices	0.67 [0.20-1.00]	**0 [0-3]**	0.56 [0.33-0.80]

Weighting

The median [IQR] of the TPR, FPC and F1 score for the base CNN, fine-tuned CNN without weights and fine-tuned CNN with weights are presented in Table 3. The fine-tuning was done for 100 iterations and included all metastases slices. Putting more weight on the missed liver metastases, leads to more detected metastases, but at the cost of more false positive objects.

Qualitative results

Some visual examples of lesion detection by the base CNN and by the patient-specific

CNN are given in Figure 3. The patient-specific CNN is fine-tuned in 100 iterations, including all slices with metastases present, and without weights. After patient-specific fine-tuning the median TPR increases from 0.67 to 0.85, the median FPC decreases from 3 to 2, and the F1 score increases from 0.49 to 0.57.

Table 3 Median [interquartile range] of the true positive rate (TPR), the false positive count (FPC) and the F1 score of the liver metastases detection, for weighting the true positives, false negatives and false positives during the patient-specific fine-tuning (FT). The best results are printed in bold. An asterix indicates a significant difference with results of the 'Base CNN' (Wilcoxon signed rank test with p<0.01).

	TPR	FPC	F1 score
Base CNN	0.67 [0.32-1.00]	3 [0-7]	0.49 [0.21-0.67]
FT CNN	0.85 [0.44-1.00]	**2 [0-3]**	**0.57 [0.40-0.67]**
Weighted FT CNN	**1.00 [0.62-1.00]**	6 [3-14]*	0.42 [0.29-0.57]

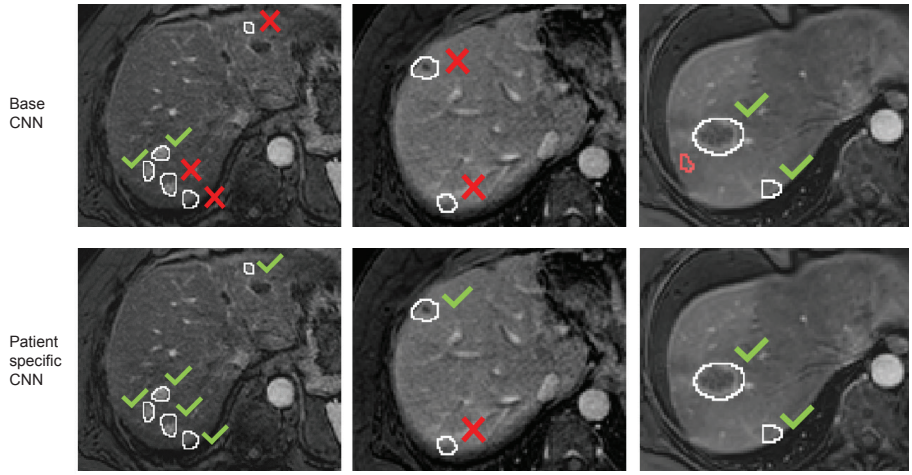


Figure 3 Examples of the detection results of the base CNN and the patient-specific CNN for three different patients. White outline = manual annotation, red outline = false positive object, green check = detected metastasis, red cross = missed metastasis.

Metastases smaller than 1.0 cm^3 are more often missed by the base CNN than larger metastases, an example of this can be seen in the first column of Figure 3. The TPR for the small metastases increases from 0.13 to 0.26 after patient-specific fine-tuning, while the TPR of the large metastases only increases slightly, from 0.81 to 0.83. The increase in TPR is thus mainly due to the higher TPR of small metastases.

Uncertainty of CNN

The softmax probabilities of the detected metastases by the base CNN have a mean SD of 0.203 (±0.078). The mean SD can be considered a measure of the uncertainty of detection. The patient-specific CNN has a mean of 0.158 (±0.070), which is significantly lower than the mean SDs of the base CNN (p=0.003, paired Student's t-test). The patient-specific CNN does not only detect more metastases, it is also more certain about the detected metastases.

WMH segmentation

Number of iterations

All four options (50, 100, 500, or 100 iterations) for the duration of the fine-tuning, give similar results. The Dice score significantly increases (Paired Student's t-test with $p<0.01$) and the AVD decreases after fine-tuning in comparison with the results of the base CNN. The base CNN has a mean DSC of 0.82 and a mean AVD of 20.7%. The fine-tuned CNNs have a mean DSC around 0.87 and a mean AVD around 10%. Using only 50 iterations gives a slightly lower Dice score (0.85) than the higher number of iterations. The number of iterations is set to 100 for the rest of the experiments.

Number of slices

The mean (± SD) of the Dice score and the AVD for different numbers of slices are presented in Table 4. The slice selection method selected the same slices when either all brain slices or all WMH slices were considered for selection. The selected slices originated mostly from the area around the ventricles and the regions with large WMH.

Including one slice already alters the Dice score and the AVD significantly. Meanwhile, further inclusion of all lesion or brain slices gives only a slightly better result. It is therefore possible to only annotate the one or two selected slices and improve the segmentation using these slices in the fine-tuning process.

Weighting

In Table 5 the mean (± SD) of the Dice score and the AVD are presented for the base CNN, fine-tuned CNN without weights and fine-tuned CNN with weights. Visual inspection of the results showed that weighting the true positive, false positive, and

false negative pixels makes the CNN more prone to label a pixel as lesion, resulting in oversegmentation. For the WMH lesion segmentation task weighting the pixels did not result in a better segmentation.

Table 4 Mean (±SD) of the Dice score and absolute volume difference (AVD) of the WMH segmentation for a varying number of slices for fine-tuning. The best results are printed in bold. A cross indicates a significant difference with results of the 'Base CNN' (Paired Student's t-test with p<0.01).

	Dice score	AVD (%)
Base CNN	0.82 ± 0.05	20.7 ± 13.5
1 slice	0.86 ± 0.04†	11.2 ± 8.2
2 slices	0.86 ± 0.04†	10.7 ± 7.9
All lesion slices	**0.87 ± 0.04**†	10.7 ± 7.3
All brain slices	**0.87 ± 0.04**†	**10.2 ± 6.9**†

Table 5 Mean (±SD) of the Dice score and absolute volume difference (AVD) of the WMH segmentation for weighting the true positives, false negatives and false positives during the patient-specific fine-tuning (FT). The best results are printed in bold. A cross indicates a significant difference with results of the 'Base CNN' (Paired Student's t-test with p<0.01).

	Dice score	AVD (%)
Base CNN	0.82 ± 0.05	20.7 ± 13.5
FT CNN	**0.87 ± 0.04**†	**10.7 ± 7.3**
Weighted FT CNN	0.86 ± 0.05†	11.8 ± 10.5

Qualitative results

Some visual examples of the lesion segmentation on the follow-up scan by the base CNN and the patient-specific CNN are given in Figure 4. The patient-specific CNN is fine-tuned in 100 iterations, including all slices with WMH lesions present, and without weights. After patient-specific fine-tuning the average Dice score increases from 0.82 to 0.87, the AVD decreases from 20.7% to 10.7%.

Other metrics of evaluation in the WMH challenge are the modified Hausdorff distance, TPR, and F1 score[15]. After patient-specific fine-tuning the Hausdorff distance decreases from 2.40 mm to 2.01 mm, the TPR decreases from 0.84 to 0.63, and the F1 score remains 0.72. The TPR decreases, because the fine-tuned CNN misses lesions with the size of only a few pixels. However, the other metrics improve due to the decrease in false positives and better volume segmentation.

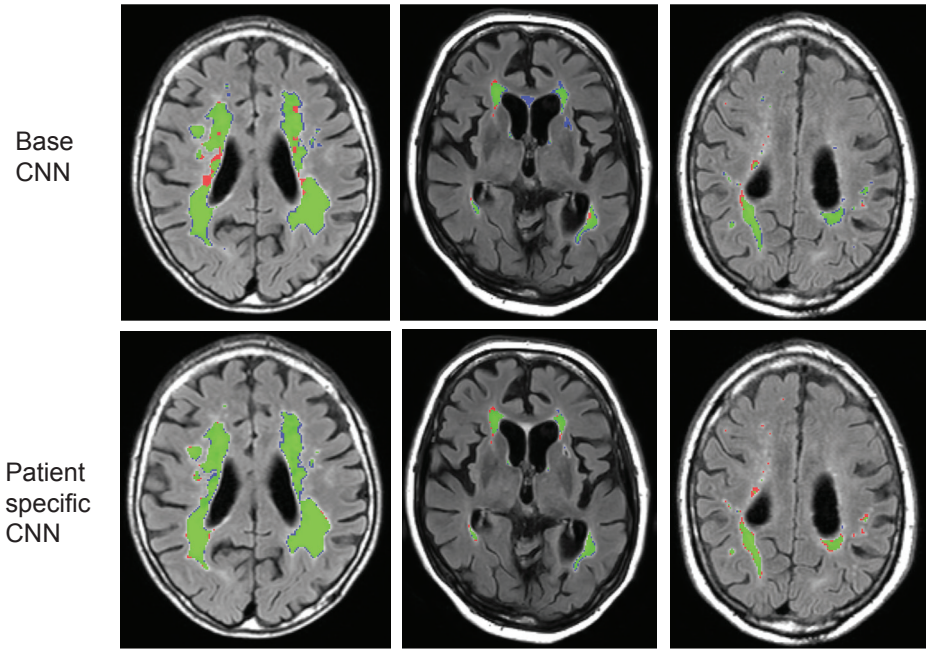


Figure 4 Examples of the follow-up scan with the segmentation results of the base CNN and the patient-specific CNN for three different patients. Green = true positive pixels, red = false negative pixels, and blue = false positive pixels.

The majority of the incorrectly labelled pixels by the base CNN are either small false positive regions (e.g. see Figure 4, second column) or false negative pixels as part of larger lesions (e.g. see Figure 4, first column). The patient-specific CNN learns to label these pixels correctly, resulting in a higher Dice score and lower AVD and thus providing a better WMH segmentation.

Smaller lesions are harder to segment than larger lesions[15] and we noticed that the patient-specific CNN missed more smaller lesions (<0.01 cm^3) than the base CNN in the segmentation, but at the same time also segmented fewer false positive pixels. The lesions and (noisy) false positive pixels labelled by the base CNN have a similar appearance, resulting in either labelling them all as lesion or all as background. The fine-tuning procedure seems to put more emphasis on reducing the smallest false positives, at the cost of small false negatives. Figure 4, column 3 shows an example of these mislabeled pixels representing small lesions.

Uncertainty of CNN

The softmax probabilities of the base CNN have a mean maximum SD of 0.398

(±0.025). The patient-specific CNN has a mean maximum SD of 0.328 (±0.038), which is significantly lower than the maximum SDs of the base CNN ($p<0.001$, paired Student's t-test). In Figure 5, an example is given of the mapped SD of the probabilities. The patient-specific CNN is more certain about the labels given, especially within large WMH.

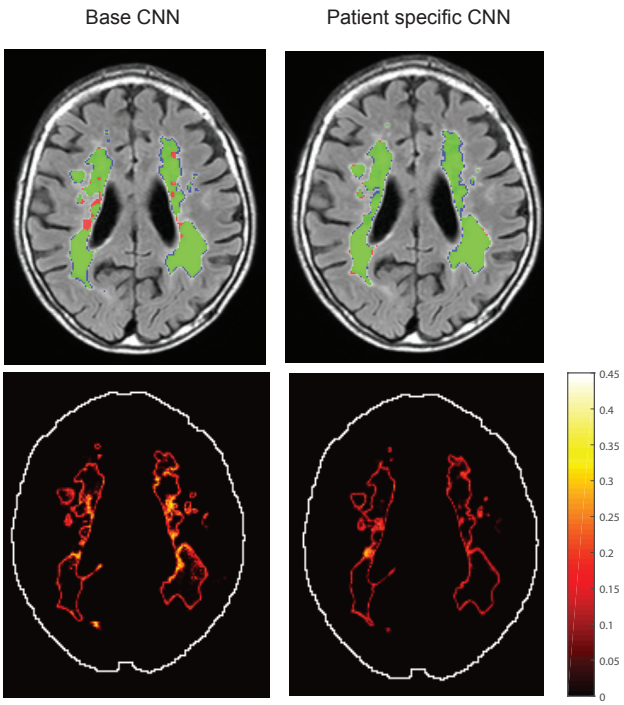


Figure 5 An example of the uncertainty (SD of softmax probability) of the base CNN and the patient-specific CNN. A high SD means the CNN is uncertain about its decision.

Discussion

Patient-specific fine-tuning of a CNN with previously acquired MR images yields improvements for quantification of lesions on subsequent MR images. The true positive rate for the detection of liver metastases and the Dice score for the WMH segmentation both increase after patient-specific fine-tuning. Additionally, the certainty of the patient-specific CNN is higher than the certainty of the base CNN.

This study shows the potential of the patient-specific fine-tuning step to improve lesion quantification results when follow-up imaging assessment is performed. Previously acquired scans of a patient are proven to hold valuable information to

refine a CNN accordingly, resulting in a better performance. The patient-specific fine-tuning step increases the lesion quantification effectiveness for two applications and we expect other applications to benefit from this fine-tuning step as well.

The loss function used is not tailored towards the detection of liver metastases but rather to the segmentation of the metastases. In case of large over- or undersegmentation on the baseline scan, it is most cost-effective for the CNN to correct those areas. Large false positive regions are caused by tissue appearances the base CNN is not familiar with, such as radiofrequency ablation (RFA) regions in the liver from earlier treatments. The patient-specific CNN will learn to correctly label the majority of the pixels in these regions. However, after patient-specific fine-tuning pixels with a similar appearance, but of the other class, can be labeled incorrectly. This is an overshoot on the follow-up scan, where the patient-specific CNN corrects its results too much. False negatives on the baseline then might cause some false positives on the follow-up due to a similar appearance and vice versa. As a consequence, more objects are detected incorrectly or more are missed, respectively. This might be overcome by tailoring the loss function towards the detection task, i.e. calculate the loss function on an object base instead of pixel base.

The patient-specific fine-tuning method assumes that the baseline and follow-up scan share features that describe the lesions and healthy tissue. However, there can be differences between the baseline scan and the follow-up scan due to (ongoing) treatment and image quality variations. The patients in the brain data set did not undergo any treatment that would be expected to influence WMH features, while most patients in the liver data set did. Therefore, the liver data set shows visually more differences between the consecutive MRI scans than the brain data set. Variation in consecutive images increases the risk of overfitting when the CNN is fine-tuned for many iterations. In addition, fine-tuning for more than 100 iterations did not improve the results compared to fine-tuning for 100 iterations.

Regarding the number of slices to include in the fine-tuning step, both applications yielded the best results when including all slices with lesions present. However, for the WMH segmentation including one or two slices gave similar results. The WMH have similar features throughout the slices and the consecutive MRI exams. This resemblance makes it possible to fine-tune the CNN with only one or two slices, reducing the time and effort invested in annotating the slices. The liver metastases have different features throughout the slices and MRI exams, requiring to fine-tune the CNN with all metastases slices.

In addition, an imbalanced data set arises for the liver metastases detection when all the liver slices are included in the fine-tuning step. That is, on average only 30% of the liver slices contain liver metastases, while 63% of the brain slices contain WMH. The CNN will then learn to distinguish different types of background instead of the appearance of liver metastases, leading to fewer false positives and unfortunately also fewer true positives. In case of an imbalanced data set, a full lesion annotation is necessary to include only the slices with lesions present and gain improvement with patient-specific fine-tuning. In other cases annotations of a selection of slices might be sufficient to gain improvement on lesion quantification.

Moreover, the difference in the percentage of slices containing lesions leads to a considerable difference in the chance that a slice with lesions is selected. This resulted in the selection of slices without liver metastases and the unsuccessful fine-tuning of the CNN. Since brain MRI contain more slices with WMH, the chance to automatically select slices with WMH present is higher, resulting in a successful fine-tuning of the CNN.

Weighting the true positives and false positives makes the detection CNN more inclined to label a pixel as lesion. This results in more detected lesions, but also more false positives in comparison with the non-weighted patient-specific CNN. Considering the aim of the detection one might allow more false positive objects in order for more detected lesions. However for the WMH segmentation, the weighting of the false positives, false negatives, and true positive pixels did not result in a better segmentation.

The information about the location of the lesions found in the baseline scan could be a valuable information for the detection method. At the moment only the appearances of the lesion and non-lesion tissue are learned. Adding the spatial information of former lesions could aid the method in finding formerly existing lesions. Image registration could be used to detect these lesions in follow-up scans. However such a method should also take into account that more lesions can develop or that lesions can disappear due to treatment.

Conclusion

In conclusion, patient-specific fine-tuning of a CNN for the quantification of lesions is a viable option to enhance the performance of the method using previously acquired data of the same patient. The CNN is fine-tuned towards the features specific for that patient resulting in a better performance of the CNN. It is important that slices with

lesions that represent the features of all lesions are included in the patient-specific fine-tuning step, as well as avoiding imbalanced classes in these slices. In doing so, the patient-specific CNN detected more liver metastases than the base CNN, with the true positive rate increasing from 0.67 to 0.85. The patient-specific CNN segments the WMH better than the base CNN with the Dice score increasing from 0.82 to 0.87.

References

1. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

2. Suzuki, K. Overview of deep learning in medical imaging. *Radiol. Phys. Technol.* **10**, 257–273 (2017).

3. Anwar, S. M. *et al.* Medical image analysis using convolutional neural networks: A review. *J. Med. Syst.* **42**, 1–13 (2018).

4. Sahiner, B. *et al.* Deep learning in medical imaging and radiation therapy. *Med. Phys.* **46**, e1–e36 (2019).

5. Mazurowski, M. A., Buda, M., Saha, A. & Bashir, M. R. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *J. Magn. Reson. Imaging* **49**, 939–954 (2018).

6. Dhara, A. K. *et al.* Segmentation of Post-operative Glioblastoma in MRI by U-Net with Patient-Specific Interactive Refinement. in *Crimi A., Bakas S., Kuijf H., Keyvan F., Reyes M., van Walsum T. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science* **11383**, 115–122 (Springer, Cham, 2019).

7. Laves, M., Bicker, J., Kahrs, L. A. & Ortmaier, T. A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *Int. J. Comput. Assist. Radiol. Surg.* **14**, 483–492 (2019).

8. Hermessi, H., Mourali, O. & Zagrouba, E. Transfer learning with multiple convolutional neural networks for soft tissue sarcoma MRI classification. in *Proc. SPIE 11041, Eleventh International Conference on Machine Vision (ICMV 2018)* (2018). doi:10.1117/12.2522765

9. Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans. Med. Imaging* **35**, 1299–1312 (2016).

10. Wang, G. *et al.* Interactive medical image segmentation using deep learning with image-specific fine-tuning. *IEEE Trans. Med. Imaging* **37**, 1562–1573 (2018).

11. Bredell, G., Tanner, C. & Konukoglu, E. Iterative Interaction Training for Segmentation Editing Networks. in *Shi Y., Suk HI., Liu M. (eds) Machine Learning in Medical Imaging. MLMI 2018. Lecture Notes in Computer Science, vol 11046* 363–370 (Springer International Publishing, 2018). doi:10.1007/978-3-030-00919-9

12. Jansen, M. J. A. *et al.* Evaluation of motion correction for clinical dynamic contrast enhanced MRI of the liver. *Phys. Med. Biol.* **62**, 7556–7568 (2017).

13. Jansen, M. J. A. *et al.* Liver segmentation and metastases detection in MR images using convolutional neural networks. *J. Med. Imaging* conditionally accepted (2019).

14. Aalten, P. *et al.* The Dutch Parelsnoer Institute - Neurodegenerative diseases; methods, design and baseline results. *BMC Neurol.* **14**, 1–8 (2014).

15. Kuijf, H. J. *et al.* Standardized Assessment of Automatic Segmentation of White

Matter Hyperintensities; Results of the WMH Segmentation Challenge. *IEEE Trans. Med. Imaging* (2019). doi:10.1109/TMI.2019.2905770

16. Boomsma, M. J. F. *et al.* Vascular cognitive impairment in a memory clinic population: Rationale and design of the "Utrecht-Amsterdam Clinical Features and Prognosis in Vascular Cognitive Impairment" (TRACE-VCI) Study. *JMIR Res. Protoc.* **6**, (2017).

17. Penny, W., Friston, K., Ashburner, J., Kiebel, S. & Nichols, T. *Statistical parametric mapping: The analysis of functional brain images*. (Academic Press, 2007).

18. Yu, F. & Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. in *ICLR* (2016). doi:10.16373/j.cnki.ahr.150049

19. Prins, N. D. & Scheltens, P. White matter hyperintensities, cognitive impairment and dementia: an update. *Nat. Rev. Neurol.* **11**, 157–165 (2015).

20. Wardlaw, J. M. *et al.* Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* **12**, 822–838 (2013).

21. Au, R. *et al.* Association of White Matter Hyperintensity Volume With Decreased Cognitive Functioning. *Arch Neurol.* **63**, 246–250 (2006).

22. Wolters, F. J. *et al.* Cerebral perfusion and the risk of dementia. *Circulation* **136**, 719–728 (2017).

23. Haley, A. P. *et al.* Subjective cognitive complaints relate to white matter hyperintensities and future cognitive decline in patients with cardiovascular disease. *Am. J. Geriatr. Psychiatry* **17**, 976–985 (2009).

24. Sander, J., de Vos, B. D., Wolterink, J. M. & Isgum, I. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. in *Proc. SPIE 10949, Medical Imaging 2019: Image Processing* 1094919 (2019). doi:10.1117/12.2511699

25. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation : Representing model uncertainty in deep learning. in *International Conference on Machine Learning (ICML)* **2016**, 1050–1059 (2016).

6

Automatic classification of focal liver lesions based on MRI and risk factors

Based on:
M.J.A. Jansen, H.J. Kuijf, W.B. Veldhuis, F.J. Wessels, M.A. Viergever, and J.P.W. Pluim, "Automatic classification of focal liver lesions based on MRI and risk factors", *PLOS ONE* 2019, 14(5): e0217053

Abstract

Accurate classification of focal liver lesions is an important part of liver disease diagnostics. In clinical practice, the lesion type is often determined from the abdominal MR examination, which includes T2-weighted and dynamic contrast enhanced (DCE) MR images. To date, only T2-weighted images are exploited for automatic classification of focal liver lesions. In this study additional MR sequences and risk factors are used for automatic classification to improve the results and to make a step forward to a clinically useful aid for radiologists.

Clinical MRI data sets of 95 patients with in total 125 benign lesions (40 adenomas, 29 cysts and 56 hemangiomas) and 88 malignant lesions (30 hepatocellular carcinomas (HCC) and 58 metastases) were included in this study. Contrast curve, gray level histogram, and gray level co-occurrence matrix texture features were extracted from the DCE-MR and T2-weighted images. In addition, risk factors including the presence of steatosis, cirrhosis, and a known primary tumor were used as features. Fifty features with the highest ANOVA F-score were selected and fed to an extremely randomized trees classifier. The classifier evaluation was performed using the leave-one-out principle and receiver operating characteristic (ROC) curve analysis.

The overall accuracy for the classification of the five major focal liver lesion types is 0.77. The sensitivity/specificity is 0.80/0.78, 0.93/0.93, 0.84/0.82, 0.73/0.56, and 0.62/0.77 for adenoma, cyst, hemangioma, HCC, and metastasis, respectively.

The proposed classification system using features derived from clinical DCE-MR and T2-weighted images, with additional risk factors is able to differentiate five common types of lesions and is a step forward to a clinically useful aid for focal liver lesion diagnosis.

Introduction

Proper differentiation between benign and malignant liver lesions is relevant to avoid unnecessary biopsies. Additionally, characterization and classification of focal liver lesions is of great importance for adequate treatment. An MRI examination of the liver, including a dynamic contrast enhanced (DCE) series, is an effective tool for detection and characterization of liver lesions by radiologists, thanks to the good soft tissue contrast. For example, hyper intensity on T2-weighted MR images, enhancement patterns on DCE-MR images and the presence of washout of contrast agent are features used in clinical practice to differentiate focal liver lesions[1,2].

Furthermore, clinical meta information and risk factors, such as symptoms, age, sex, and a known primary tumor, help in decision making[3,4]. Nonetheless, it can still be difficult and time consuming to distinguish lesion types, because of the broad range of lesion appearances on MRI.[5] An automatic classification system could aid radiologists in this task.

Some efforts have been made to automatically classify focal liver lesions using T2-weighted MR images to aid radiologists. Mayerhoefer et al. (2010)[6] applied texture-based classification on T2-weighted and T1-weighted MR images, to distinguish cysts and hemangiomas. This classification system reached accuracy rates of 0.77-0.84 for T1-weighted images and 0.75-0.88 for T2-weighted images, for differentiating two benign lesion classes. Another classification system based on features derived from T2-weighted MR images, developed by Gatos et al. (2017)[7], differentiates between benign lesions, hepatocellular carcinoma (HCC) within a cirrhotic liver, and metastases within a noncirrhotic liver. This system reached an overall accuracy of 0.90. However, it was not designed to distinguish different types of benign lesions. Both methods applied features derived from T2-weighted images, but did not include risk factors or features derived from DCE-MR images, while these features may supply additional information for a classification system to differentiate more lesion types[4,8].

This chapter proposes a classification method that aims to differentiate between five lesion classes: adenomas, cysts, hemangiomas, HCCs and metastases by exploiting features derived from DCE-MR images with an extracellular contrast agent, as well as features from T2-weighted images. Risk factors for adenoma, HCC and metastasis were also taken into account as features.

Materials and methods

Data

The study comprises MRI data of patients with suspicion of liver lesions from the University Medical Center Utrecht, The Netherlands, acquired between February 2015 and February 2017. The UMCU Medical Ethical Committee has reviewed this study and informed consent was waived due to its retrospective nature. Patients without lesions, with lesions other than the common adenomas, cysts, hemangiomas, HCCs or metastases, or with atypical lesions were excluded. Focal nodular hyperplasias (FNHs) were not included, because there was an insufficient number of FNHs for training a classifier. Also, liver lesions with a diameter of less than 5 mm were

excluded from this study. Up to four lesions per patient were included. In order to balance the classes, the data sets from patients with liver metastases acquired between February 2015 and February 2016, were excluded. In total, 95 patient data sets with 125 benign lesions (40 adenomas, 29 cysts and 56 hemangiomas) and 88 malignant lesions (30 HCCs and 58 metastases) were included in this study. The origin of the primary tumor of the metastases was widespread, including: 8 breast carcinomas, 47 gastrointestinal carcinomas (23 colorectal carcinomas, 18 neuroendocrine carcinomas, 4 esophagus carcinomas, 1 HCC metastasis, and 1 adenocarcinoma), and 3 have another primary tumor origin. The metastases were therefore of both hypervascular (25) and hypovascular (32) type[1].

MR imaging

All 95 patients had an MRI examination on a 1.5 T scanner (Philips) with the clinical focus on the liver, including a DCE-MRI and a T2-weighted scan. Examples of lesions in DCE-MRI and T2-weighted images are shown in Figure 1. The DCE-MR images were acquired in six breath holds with one to five 3D images per breath hold. The DCE-MRI series was acquired using a clinical protocol with the following parameters: TE: 2.143 ms; TR: 4.524 ms; flip angle: 10 degrees. After acquiring the first image, gadobutrol (0.1 ml/kg Gadovist of 1.0 mmol/ml at 1 ml/s) was administered at once, followed by 25 ml saline solution at 1 ml/s. In total, 16 3D images per patient were acquired with 100-119 slices and matrix sizes ranging from 256 × 256 to 288 × 288. Voxel size was 1.543 mm × 1.543 mm × 2 mm. The T2-weighted images were acquired during free breathing using a TSE protocol with the following parameters: TE: 80 ms; TR: 756 ms; flip angle: 90 degrees. Matrix sizes ranged from 448 × 448 to 512 × 512, with 25-30 slices. Voxel size was 0.938 mm × 0.938 mm × 8 mm. Eight patients underwent a slightly modified T2-weighted acquisition due to a change in the clinical scanning protocol, with: TE: 102 ms; TR: 2945 ms. The slice thickness became 5 mm, resulting in 41 slices for the same FOV. Two other patients were scanned with a fat-suppressed T2-weighted SPAIR protocol, in which the voxel size was 0.893 mm × 0.893 mm × 5 mm and the matrix size was 448 × 448 with 46 slices. All scans were acquired in axial direction.

The DCE-MR series were corrected for motion using a PCA-based groupwise registration algorithm[9] as described in Jansen et al. (2017)[10]. The intensities in the DCE-MR series were linearly remapped between 0 (background) and 1 (contrast agent peak in the aorta) for standardization. For the T2-weighted images the original intensities were kept, because the highest intensity is dependent on the type of lesion, i.e. cysts can have a higher intensity than the gallbladder or the spinal fluid.

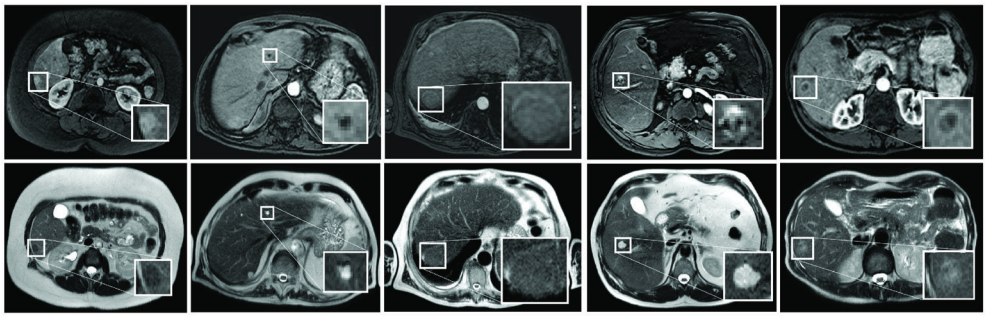


Figure 1 Examples of focal liver lesions. From left to right livers with an adenoma, cyst, HCC, hemangioma, and a metastasis from colorectal carcinoma origin are shown. The top row shows the arterial phases of the DCE-MRI and the bottom row the T2-weighted images. A zoom-in of the lesions is inserted.

All lesions were manually segmented on a single slice in both the motion-corrected DCE-MR series and the T2-weighted images by a researcher and were verified by an independent expert. Malignant lesions were additionally proven by pathology or follow-up. However, for lesions classified by an expert radiologist as benign, a follow-up was not available. The area of the lesion segmentations ranged from 0.2 to 53.0 cm^2 in the 2D slice, with a median of 1.8 cm^2. Furthermore, a small part, on average 3.5 cm^2, of healthy parenchyma was manually segmented in the DCE-MR series for feature calculation.

Feature extraction

In total, 164 features were extracted from the DCE-MR and T2-weighted MR images within the lesion, including contrast curve features, gray level histogram features, and GLCM texture features. Part of the features were chosen based on characteristics used by radiologists during visual rating. The other part of the feature categories were already proven to be of value in the classification of liver lesions [6,7] or breast lesions with DCE-MR imaging [11,12]. Risk factors, as the presence of steatosis, cirrhosis, and the presence or absence of a known primary tumor somewhere in the body, were also used as features. An overview of all features is given in Table 1. Below we will explain how the contrast curves, images, and histograms are defined, on which the features are calculated.

Table 1 Features derived from DCE-MR and T2-weighted images and risk factors.

Categories	Features	Calculated on:
Contrast curve features	Maximum enhancement, time to peak (TTP), uptake rate, washout rate [11], area under the curve (AUC), average plateau, early-to-late signal enhancement ratio (SER) and time of arrival[12]	TIC, CEC, CE lesion-parenchyma-ratio, TIC lesion-parenchyma ratio
Gray level histogram features	Mean, standard deviation, skewness, kurtosis and the 10^{th} and 90^{th} percentile of the intensities	Pre-contrast image, TTP image, and late enhancement phase of the DCE-MR series, T2-weighted image, TTP feature map, radial gradient histogram of late arterial enhancement phase[13], ring enhancement histogram of portal-venous phase.
Texture features (gray level co-occurrence matrix (GLCM) features)	Angular second moment, contrast, correlation, sum of squares variance, homogeneity, sum average, sum variance, entropy, sum entropy, difference variance, difference entropy, IMC1 and IMC2; calculated from the summed gray level co-occurrence matrix (GLCM) in 4 directions, 0˚, 45˚, 90˚ and 135˚, with an offset of 1 pixel	Pre-contrast image, TTP image, and late enhancement phase of the of DCE-MR series, variance in all DCE-MR images of the series, T2-weighted image, TTP feature map
Risk factors and other	Presence of steatosis, cirrhosis, and primary tumor in the body, area of lesion	

TIC = time intensity curve, CEC = contrast enhanced curve, CE = contrast enhanced.

Contrast curves

The contrast curve features were obtained from the lesion time intensity curve (TIC) and three curves derived from it. To smooth the TIC, the motion-corrected DCE-MR series were first filtered with the TIPS bilateral filter[14]. The TIC was obtained for each pixel in the lesion mask and in the healthy parenchyma mask. The three other

contrast curves are the contrast enhancement curve (CEC), the contrast enhancement (CE) lesion-parenchyma ratio, and the TIC lesion-parenchyma ratio. The CEC was calculated by dividing the lesion TIC by the intensity of the pixel in the pre-contrast image. The CE lesion-parenchyma ratio was obtained by dividing the average CEC of the lesion ROI time point-wise by the average CEC of the parenchyma ROI, which was calculated in a similar manner as the CEC of the lesion. The TIC lesion-parenchyma ratio was obtained by dividing the average TIC of the lesion ROI time point-wise by the average TIC of the parenchyma ROI. The average curves were obtained by taking the average for each time point within the lesion and were used to calculate the features, as given in Table 1. For the CEC and the CE lesion-parenchyma ratio curves, the standard deviation of the feature values within the lesion was included as feature.

Images and histograms

Gray level histogram features and GLCM texture features were derived from the pre-contrast image, the time-to-peak (TTP) image and late arterial enhancement phase image of the DCE-MR series, as well as from the T2-weighted image and the TTP feature map. The variance of the GLCM texture features throughout the DCE-MR series was also included. The TTP image was defined as the image of the DCE-MR series in which the maximum enhancement of the lesion in the average CE curve was reached. The TTP feature map was defined as the TTP value in seconds obtained per voxel. Two examples of TTP feature maps are shown in Figure 2.

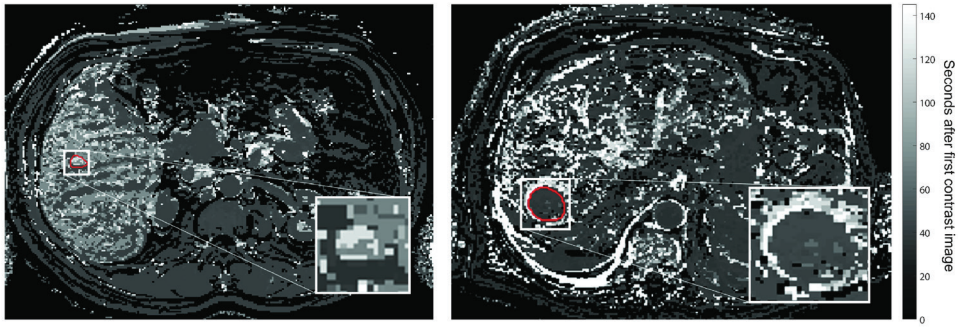


Figure 2 Time-to-peak feature maps of a liver with a hemangioma (left) and a HCC (right). The lesions correspond with the lesions in Figure 1. The red contours show the lesion segmentations.

Additionally, histogram features were calculated from the radial gradient histogram and the ring enhancement histogram. The radial gradient histogram was obtained from the late arterial enhancement phase, multiplying the difference in intensity between each voxel value and the lesion's central voxel value with the distance between this voxel and the central voxel. The values are then normalized between 0 and 1. The

features derived from this histogram provide information about the circularity of the enhancement within the lesion. Mean values around 1 indicate a homogenous spherical pattern in the lesion[13]. The ring enhancement histogram was obtained by calculating the normalized distances between the center of the lesion and the pixel with the maximum first derivative of the portal-venous phase intensities, in all radial directions. The first derivatives were calculated between the central voxel and the edges of the lesion segmentation mask plus two extra voxels. The portal-venous phase would show the ring enhancement if present. A mean ring enhancement of 1 indicates ring enhancement and a high SD indicates heterogeneous enhancement.

Risk factor and other features

If the presence of steatosis was stated in the clinical report, the corresponding feature was assigned the value one, in absence the value zero. The same was done for the presence of cirrhosis and for the presence of a primary tumor. The area of the lesion was equal to the number of pixels within the lesion mask times the pixel area.

Classification

An extremely randomized trees classifier[15] was chosen for classification, as it is a non-parametric, robust classifier that assembles decision trees using all available training samples at a low computational cost. At each node of a tree, a random set of thresholds is applied to a random subset of the features. The feature with the best threshold is selected for that node. In this way the most informative features are mostly used in the nodes of the trees. The usage of the extremely randomized tree classifier makes the need for the optimal feature selection less strict. To avoid overfitting a minimum of five samples per leaf was set, feature selection was performed, and the number of features offered to each node in the classifier was equal to the square root of the total number of features.

The classification system was implemented using scikit-learn[16]. A built-in function for feature selection applying the ANOVA F-score is used to exclude the features without any information regarding the classification of the liver lesions.

Experiments

The experiments are split into two parts. First we will demonstrate the impact of adding features from DCE-MR images and risk factors to the T2-weighted MR images features by training the extremely randomized trees classifier on different

feature sets, in a similar way as was reported by Jansen et al. (2018)[17]. Secondly, we focus on the optimization and classification of the lesion using all features.

In this part of the study, feature set A) contained only the features from T2-weighted MR images, feature set B) contained only the features from T2-weighted MR images and the risk factor features, feature set C) contained the features from T2-weighted MR images and DCE-MR images, and feature set D) contained all the available features. For a fair comparison, only the 19 features with the highest ANOVA F-score were selected, because the smallest feature set (A) only comprises 19 features. The number of features offered to each node in the classifier was equal to four, the square root of the total number of features; in total 700 trees were built. Sensitivity, specificity, and accuracy are calculated based on a leave-one-patient-out cross-validation. The leave-one-patient-out cross-validation was selected to be able to have enough training samples for generalization of the large variations in the data. The results are compared for the classification of the five lesion classes.

For the second part of this study, the optimal number of features and trees for the classifier were determined during an optimization process using cross-validation with the number features ranging from 10 to 164, in steps of 10, and the number of trees ranging from 100 to 800, in steps of 50. In this way, features that do not hold any information for the classification of these liver lesions are excluded. For this classification problem 50 features and 700 trees were optimal and used for the next experiments. The number of trees had only a small impact on the classification results, while the number of features should be chosen more carefully. Too many features could lead to a large amount of non-informative trees and overfitting.[18]

Feature selection was performed on all features to remove redundant features and speed up the classification. The 50 features with the highest ANOVA F-score were selected for classification. The number of features offered to each node in the classifier was equal to seven, the square root of the total number of features; in total 700 trees were built. The classes were balanced per sample.

The classification system was evaluated by calculating the sensitivity, specificity and the overall accuracy based on a leave-one-patient-out principle. Feature selection was repeated during the leave-one-patient-out evaluation to avoid a positive bias. Additionally, for each of the five classes, a one-versus-others accuracy was calculated by splitting the confusion matrix in one lesion class versus the other lesion classes. The receiver operating characteristic (ROC) curve is also obtained for each lesion class, showing the strength of the classifier for individual lesion classes. The ROC curve

for a particular lesion class was calculated by adding the probabilities of the other four lesion classes as other-class. The optimal cut-off values and the corresponding true positive rate (TPR), false positive rate (FPR) and false negative rate (FNR) were obtained for each lesion type using these ROC curves.

The overall accuracy, sensitivity and specificity for differentiating benign and malignant lesions was calculated by separating benign and malignant lesions in the confusion matrix. Furthermore, the ROC curve of benign-versus-malignant lesions was obtained by plotting the added probabilities of the benign lesions versus those of the malignant lesions. Using this ROC curve the optimal cut-off value and the corresponding TPR, FPR and FNR were obtained.

Results

The results of the first part of the experiments are presented in Table 2, showing the classification results for the different feature sets. The results show that in general the sensitivity, specificity and overall accuracy increase, when DCE-MR image features, risk factor features, or both are added to the T2-weighted MR image features. All feature sets B, C, and D improve significantly over feature set A, using a Chi-Square McNemar test. Therefore, all the features were used to train an extremely randomized trees classifier with an optimal number of features.

Table 2 Classification results for the four different feature sets (sensitivity/specificity), with 19 features selected.

	A	B	C	D
Adenoma	0.43 / 0.50	0.68 / 0.75	0.65 / 0.65	0.50/ 0.64
Cyst	0.76 / 0.79	0.72 / 0.75	0.93 / 0.90	1.00 / 0.91
Hemangioma	0.77 / 0.75	0.73 / 0.73	0.79 / 0.81	0.75 / 0.82
HCC	0.57 / 0.36	0.73 / 0.58	0.63 / 0.56	0.77 / 0.55
Metastasis	0.41 / 0.51	0.60 / 0.64	0.69 / 0.73	0.66 / 0.67
Overall accuracy	**0.58**	**0.69**	**0.73**	**0.71**

A) T2-weighted MR features, B) T2-weighted MR features and risk factor features, C) T2-weighted MR and DCE-MR features, D) all features.

The remainder of this section shows the results of the second part of the experiments, based on all features (feature set D). Table 3 shows the 42 features that are selected in every leave-one-patient-out repetition. Four features that are selected in >90% of the leave-one-patient-out repetitions are also included in Table 3 and indicated with an asterisk. In addition, three features are selected in >75% of the leave-one-patient-out repetitions: the sum of squares variance of the TTP image, and the standard

deviations of the gray level histogram of the TTP feature map and the radial gradient histogram. The less frequently selected features (<50%) are: the presence of cirrhosis, the presence of a primary tumor in the body, the IMC2 of the pre-contrast image, and standard deviation of the gray level histogram of the late enhancement image. Additionally, from the TTP image the sum average, sum variance, and IMC1 were selected less frequently.

Table 3 Selected features with the highest ANOVA F-scores.

Contrast curve features	Gray level histogram features	Texture features	Risk factors and other
TTP *CEC lesion-parenchyma ratio*	Mean, 10th perc., 90th perc. *Pre-contrast image*	SSVar, SumVar*, IMC1 *Pre-contrast image*	Presence of steatosis
Max. enhancement, SER, AUC *TIC*	Mean, SD*, 10th perc., 90th perc. *TTP image*	SSVar, sum average, SumVar *Late enhancement image*	
Max. enhancement, TTP, uptake*, average plateau, AUC *TIC lesion-parenchyma ratio*	Mean, 10th perc., 90th perc. *Late enhancement image*	Correlation, SSVar, sum average, SumVar, sum entropy, DiffVar, IMC 1 and 2 *T2-weighted image*	
	Mean, SD, skewness, 10th perc., 90th perc. *T2-weighted image*	Contrast*, SSVar, SumVar, DiffVar *TTP feature map*	
	Mean, 90th perc. *TTP feature map*		

TTP = time to peak; SER = early-to-late enhancement ratio; AUC = area under the curve; SSVar = sum of squares variance; SumVar = sum variance; DiffVar = difference variance. The input image/histogram of the features is listed in italic print. The asterisks indicate a feature selected in >90% of the leave-one-patient-out repetitions.

The confusion matrix with the sensitivity and specificity per class, together with the overall accuracy is presented in Table 4. The one-versus-others accuracy is given per lesion class. The overall accuracy is 0.77.

Table 4 Confusion matrix of the five class problem, including the sensitivity, specificity and one-versus-other accuracy per lesion class.

	Adenoma	Cyst	Heman-gioma	HCC	Meta-stasis	Sens	Spec	One-vs-other accuracy
Adenoma	32	0	2	4	2	**0.80**	**0.78**	**0.92**
Cyst	0	27	2	0	0	**0.93**	**0.93**	**0.99**
Hemangioma	3	2	47	0	4	**0.84**	**0.82**	**0.91**
HCC	3	0	0	22	5	**0.73**	**0.56**	**0.88**
Metastasis	3	0	6	13	36	**0.62**	**0.77**	**0.85**
Overall accuracy: 0.77								

The rows represent the true class and the columns represent the predicted class.

The ROC curves for each of the five lesion classes are shown in Figure 3. In Table 5 the area under the ROC curve (AUC) is given, together with the optimal cut-off values for each individual lesion class and the corresponding TPR, FPR and FNR.

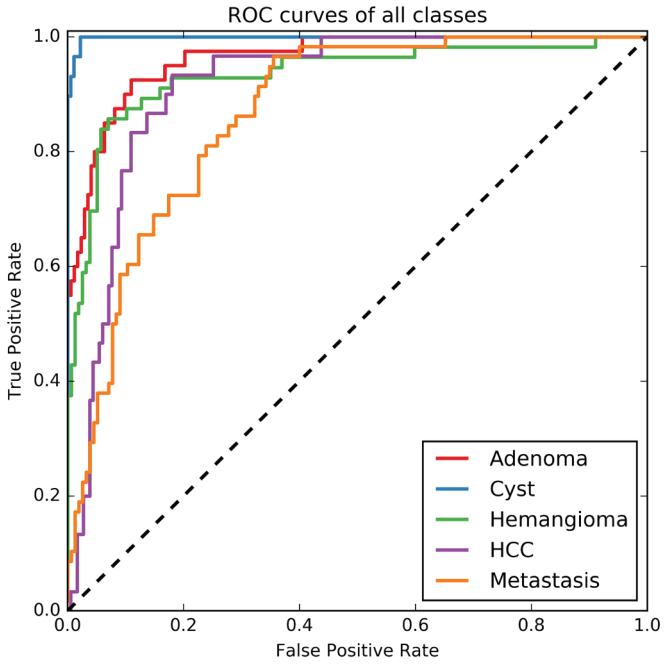


Figure 3 ROC curves of all lesion classes in a one-versus-other approach. The rest class is calculated as the outcome probabilities of the other four lesions.

Table 5 Areas under the ROC curve (AUC) for each class including the optimal cut-off value and the corresponding true positive rate (TPR), false positive rate (FPR) and false negative rate (FNR).

	AUC	Optimal cut-off value	TPR	FPR	FNR
Adenoma	0.96	0.28	0.88	0.08	0.12
Cyst	1.00	0.43	0.97	0.01	0.03
Hemangioma	0.93	0.26	0.89	0.13	0.11
HCC	0.91	0.33	0.83	0.11	0.17
Metastasis	0.88	0.26	0.81	0.24	0.19

Benign versus malignant lesions

Splitting the lesions in the confusion matrix in benign lesions (adenoma, cyst and hemangioma) and malignant lesions (HCC and metastasis), gives a sensitivity of 0.92 for benign lesions and 0.86 for malignant lesions. The specificity is 0.91 and 0.88 for benign and malignant lesions respectively and the overall accuracy is 0.90. The ROC curve for the benign-vs-malignant lesions is shown in Figure 4. The classifier gives a TPR of 0.89, a FPR of 0.20 and a FNR of 0.11 for the malignant lesions at the optimal cut-off value of 0.42. The area under the ROC curve is 0.94 for the benign-versus-malignant classification problem.

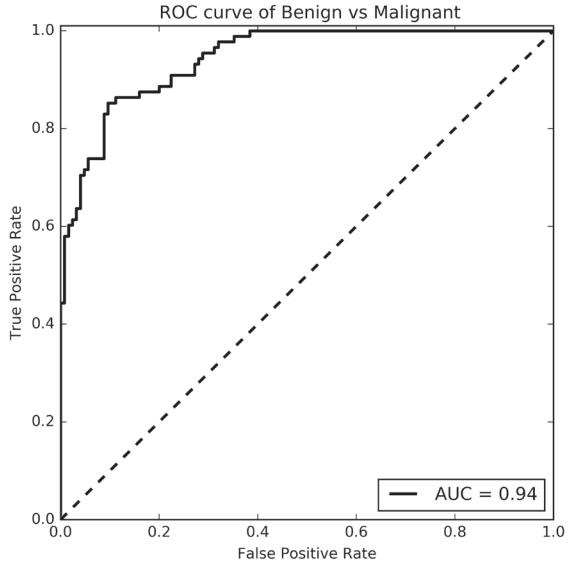


Figure 4 ROC curve of benign-versus-malignant classification problem. The area under the ROC curve (AUC) is 0.94.

Discussion

In this study, a classification system was proposed based on features derived from T2-weighted and DCE-MR images, and from risk factors. The classification system was able to differentiate five lesion types: adenoma, cyst, hemangioma, HCC, and metastasis, with an overall accuracy of 0.77. The average sensitivity [min-max] is 0.79 [0.62-0.93] and the average specificity [min-max] is 0.77 [0.56-0.93].

The classification system was challenged by the fact that the lesion types had a wide range of appearances and thus a large variance in feature values. For example, smaller HCCs and hemangiomas have a different appearance than larger ones on both DCE-MR and T2-weighted MR images[5,19]. Furthermore, there was a large variety in the origin of the primary tumor of the metastases, but all these metastases were treated as one lesion type, leading to some variety in lesion appearance. The large variance in appearances leads to overlapping feature values between lesion types. Nonetheless, the classifier was able to correctly determine the lesion type in the vast majority of cases. The unequal distribution of the origin of the primary tumor did not seem to have an effect on the classification results. When we separately analyze the three major origin types (colorectal carcinomas, neuro-endocrine tumors, and breast carcinomas), they have similar sensitivities. Only cysts have very typical features in both T2-weighted MR and DCE-MR images, and were therefore the only lesions with a small variance in feature values, resulting in the highest sensitivity and specificity.

The area under the ROC curve for each of the individual lesion types is high, ranging from 0.86 to 1.00, implying that the classifier is highly capable of distinguishing one lesion type from the others. The optimal cut-off values in these ROC curves are low in comparison to what one might expect in a two-class problem, but in this study the probabilities of the other four lesions are added instead of running the classifier again on the 2-class problem. The TPR and the FPR result from selecting the lesion type for which the probability of that lesion exceeds the optimal cut-off value as given in Table 5, instead of selecting the lesion type with the highest probability, as was done for Table 4. The assignment of a lesion type based on the optimal cut-off value results in the possibility that one lesion could have been assigned two labels. For this reason the FPR and TPR results of Table 5 do not match the sensitivity and specificity results from the confusion matrix in Table 4.

The overall accuracy for benign and malignant lesions, when splitting the confusion matrix of Table 4, is 0.90. This is higher than the five lesion types' classification, largely due to the mix up of metastases and HCC, both malignant lesions. The classifier is

thus not only able to differentiate five types of lesions, but is also able to differentiate benign from malignant lesions with an even higher accuracy, even when it is not trained for this classification problem.

Training the classifier on four different feature sets shows that both risk factor features and DCE-MR image features greatly improved the classification results, despite the fact that the number of features is suboptimal for feature set C and D. Only for hemangiomas and cysts did the addition of risk factor features not improve the classification results for the suboptimal feature set.

Unfortunately, no public data set for this application is available, but we compare the proposed method with a recent study by Gatos et al. (2017)[7] by performing the evaluation in a similar way. The study by Gatos et al. (2017) reported an overall accuracy of 0.90 for the classification of benign, HCC and metastatic focal liver lesions using only T2-weighted MR images. Our classifier was not able to obtain high accuracies with only T2-weighted MRI features, possibly because our data set included different types of lesions and has more classes than the study by Gatos et al. (2017). Nonetheless, similar one-vs-other accuracies per class were reported for benign, HCC, and metastasis classes, when exploiting features from T2-weighted MR images, DCE-MR images and risk factors.

Even though not all T2-weighted MR images in this study were acquired using the same protocol, the classifier was able to learn to correctly identify the lesions independent of small changes in the T2-weighted protocol used. This shows that the classifier is robust to slight changes in certain features.

The feature selection procedure was performed separately for every leave-one-patient-out repetition, to avoid any positive bias. Out of all repetitions, 42 features were selected in every repetition and four in >90% of the repetitions. This might indicate that the core set of 42 features are important for the classification task.

The selected features originate from all four feature categories and from DCE-MR and T2-weighted images, which underpins the importance of a wide selection of features. The optimization process for the number of features and trees in the extra trees classifier showed that the number of features had a larger impact than the number of trees. Yet, the number of trees should be sufficient for each feature to be selected in the nodes of the trees at least a few times. Having too many (redundant) features, i.e. more than 100 features, as input for the classifier will lead to overfitting and decreases the accuracy with a few percentages.

Limitations

Five common lesion types were included in this study, which does not cover the full range. Lesion types with very low occurrence in the dataset could not be included in the classification. A larger data set with a vast amount of samples per lesion type would be needed to train a classification system that is able to differentiate more lesion types with a better accuracy.

Feature calculations were done in 2D, because this reflected how the manual annotations were made and provided to the classifier. If the final method were to be applied in a clinical workflow, being able to perform classification based on a 2D lesion annotation is beneficial over requiring 3D lesion annotations. The latter are more time-consuming to create. Automatic 3D lesion segmentation, if available, would enable 3D feature calculations, which potentially further improves the performance of our proposed method.

Moreover, lesions smaller than 5 mm in diameter were excluded because texture features could not be reliably computed on lesions smaller than 3×3 pixels. Doing the calculations in 3D might overcome part of this problem.

Future work

The features in this study are based on lesion characteristics used by radiologists in clinical practice. Although the features are carefully designed, they might not be the best representation of the focal liver lesions. Learning features from the data instead of using hand-crafted features might improve the classification results. Unsupervised feature learning algorithms in combination with a supervised classifier have been used for classification of lesions in other applications.[20] For example, restricted Boltzmann machines and convolutional sparse auto-encoders have successfully been applied as feature extractors and combined with machine learning classifiers[21, 22]. In future work, the possibilities of such representation learning algorithms to automatically generate a feature set with a better representation of the liver lesions could be investigated on a larger data set.

Conclusion

The proposed classification system, based on features derived from clinical DCE-MR and T2-weighted images, as well as risk factors, is able to classify five common focal liver lesion types (adenoma, cyst, hemangioma, HCC, and metastasis) with an overall

accuracy of 0.77, and to differentiate between benign and malignant lesions with an overall accuracy of 0.90. This is a step forward to a clinically useful aid for focal liver lesion diagnosis.

References

1. Fowler, K. J., Brown, J. J. & Narra, V. R. Magnetic resonance imaging of focal liver lesions: Approach to imaging diagnosis. *Hepatology* **54**, 2227–2237 (2011).

2. Ba-ssalamah, A. *et al.* Clinical value of MRI liver-specific contrast agents: A tailored examination for a confident non-invasive diagnosis of focal liver lesions. *Eur. Radiol.* **19**, 342–357 (2009).

3. Holalkere, N. S. *et al.* Characterization of small liver lesions: Added role of MR after MDCT. *J. Comput. Assist. Tomogr.* **30**, 591–596 (2006).

4. Hamm, B. *et al.* Focal liver lesions: Nonenhanced and characterization dynamic contrast MR imaging. *Radiology* **190**, 417–423 (1994).

5. Albiin, N. MRI of focal liver lesions. *Curr. Med. Imaging Rev.* **8**, 107–116 (2012).

6. Mayerhoefer, M. E. *et al.* Texture-based classification of focal liver lesions on MRI at 3.0 Tesla : A feasibility study in cysts and hemangiomas. *J. Magn. Reson. Imaging* **32**, 352–359 (2010).

7. Gatos, I. *et al.* Focal liver lesions segmentation and classification in nonenhanced T2-weighted MRI. *Med. Phys.* **44**, 3695–3705 (2017).

8. Elsayes, K. M. *et al.* Focal hepatic lesions: Diagnostic value of enhancement pattern approach with contrast enhanced 3D gradient-echo MR imaging. *Radiographics* **25**, 1299–1320 (2005).

9. Huizinga, W. *et al.* PCA-based groupwise image registration for quantitative MRI. *Med. Image Anal.* **29**, 65–78 (2016).

10. Jansen, M. J. A. *et al.* Evaluation of motion correction for clinical dynamic contrast enhanced MRI of the liver. *Phys. Med. Biol.* **62**, 7556–7568 (2017).

11. Chen, W., Giger, M. L., Lan, L. & Bick, U. Computerized interpretation of breast MRI: Investigation of enhancement-variance dynamics. *Med. Phys.* **31**, 1076–1081 (2004).

12. Khalifa, F. *et al.* Models and methods for analyzing DCE-MRI: A review. *Med. Phys.* **41**, (2014).

13. Gilhuijs, K. G. A., Giger, M. L. & Bick, U. Computerized analysis of breast lesions in three dimensions using dynamic magnetic resonance imaging. *Med. Phys.* **25**, 1647–1654 (1998).

14. Mendrik, A. M. *et al.* TIPS bilateral noise reduction in 4D CT perfusion scans produces high-quality cerebral blood flow maps. *Phys. Med. Biol.* **56**, 3857–3872 (2011).

15. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).

16. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2012).

17. Jansen, M. J. A., Kuijf, H. J. & Pluim, J. P. W. Automatic classification of focal liver lesions based on clinical DCE-MR and T2-weighted images: A feasibility study. in *2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018* **2018**, 245–248 (2018).

18. Jain, A. K., Duin, R. P. . & Mao, J. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 4–37 (2000).

19. Matos, A. P. *et al.* Focal liver lesions: Practical magnetic resonance imaging approach. *World J. Hepatol.* **7**, 1987–2008 (2015).

20. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).

21. Van Tulder, G. & De Bruijne, M. Combining generative and discriminative representation learning for lung CT analysis with convolutional restricted Boltzmann machines. *IEEE Trans. Med. Imaging* **35**, 1262–1272 (2016).

22. Kallenberg, M. *et al.* Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Trans. Med. Imaging* **35**, 1322–1331 (2016).

7

Summary and general discussion

Summary

Lesion detection and characterization is an important step in the diagnosis and treatment planning of focal liver lesions such as hemangiomas, adenomas, and liver tumors. Image processing and image analysis using machine learning can aid radiologists performing these tasks. In this thesis several methods regarding image registration, lesion detection, and lesion classification were developed and evaluated. The methods were developed for dynamic contrast enhanced (DCE) MR images, sometimes in combination with other MR sequences.

DCE-MR images contain both structural and functional information, which makes them very useful for image processing and analysis. However, motion between the different contrast enhancement phases is introduced by inconsistent breath hold and by bowel and cardiac movements. In Chapter 2, we evaluated two image registration methods, a groupwise and a pairwise registration method. The evaluation was based on quantitative and qualitative metrics. The results showed that the groupwise registration achieved better temporal alignment with smoother spatial deformations than pairwise registration. The results provided by the groupwise registration are so satisfactory that clinical abdominal DCE-MR images are automatically registered and presented to the radiologists for analysis in their current workflow. Moreover, the groupwise registration method was used for the other studies described in this thesis.

Next, a liver segmentation method using DCE-MR images was developed. The liver segmentation was used as a first step in the liver lesion detection method, to define the region of interest. In Chapter 3, the three input configurations of DCE-MR images for a convolutional neural network (CNN) were studied, with the aim to obtain the best liver segmentation. The three configurations are I) one phase image of the DCE-MR images as input image; II) the separate phases of the DCE-MR images as input images; and III) the separate phases of the DCE-MR images as channels of one input image. We found that having the six phases of the DCE-MR images as different channels of the input image gives the most satisfying results, independent of the CNN architecture.

In Chapter 4, the liver segmentation method was more extensively evaluated and a novel liver lesion detection method was introduced. In this chapter we found that DCE-MR images alone were not sensitive enough for liver lesion detection. The inclusion of diffusion weighted (DW) MR images was proven to be a valuable addition to the detection method, since the detection rate increased with more than 50 per cent. Furthermore, it was found that concatenating the DCE-MR and DW-MR

images to form one input image was not as effective as having a dual pathway CNN, one path for each of the MR sequences.

The liver lesion detection method described in Chapter 4 is a general method and could fail if a patient presents features not known or seen before by the CNN. In Chapter 5, a patient-specific fine-tuning approach was proposed to obtain a detection method dedicated towards one patient. This approach was not only evaluated on the aforementioned detection of liver lesions, but also on segmentation of white matter hyperintensities in brain MRI. During patient-specific fine-tuning, previously acquired MR examinations were exploited to update the pre-trained CNN towards the patient-specific features. This approach improved the performance of both lesion quantification methods.

Once the lesions are detected, the type of lesion needs to be determined. In Chapter 6, a lesion classification method was proposed. The classification method was able to successfully distinguish benign from malignant lesions as well as differentiating five focal liver lesions: adenoma, cyst, hemangioma, HCC, and liver metastasis. Contrast curve, gray level histogram, and gray level co-occurrence matrix texture features were extracted from the DCE-MR and T2-weighted images. In addition, risk factors including the presence of steatosis, cirrhosis, and a known primary tumor were used as features. Fifty features with the highest ANOVA F-score were selected and fed to an extremely randomized trees classifier. The fifty selected features originated from all four feature categories and from DCE-MR and T2-weighted images, which underpins the importance of a wide selection of features.

General discussion

In clinical practice CT is still commonly used for liver imaging. Most image processing and image analysis methods for the liver and liver lesions are therefore developed for CT images[1–4]. However, liver MR imaging is becoming increasingly popular thanks to a better contrast between liver and lesion without the use of ionizing radiation. In this thesis, some methods are proposed for the detection and classification of liver lesions by exploiting MR images and dynamic contrast enhanced (DCE) MR images in particular.

Multiparametric MR imaging

DCE-MR images consist of both functional and structural information, useful for image processing and analysis. In this thesis we showed that DCE-MR images can be

applied to obtain accurate liver masks. However, for the detection and classification of liver lesions the mere use of DCE-MR images is less effective. In Chapters 4 and 6, the value of the addition of other MR sequences was investigated. Both chapters show that the use of multiple MR sequences, also known as a multiparametric approach, are preferred over the use of a single MR sequence in terms of performance.

The multiparametric image processing and analysis could be a valuable approach for liver MR images, since it combines the information from different functional and structural imaging techniques. Multiparametric MR imaging is already proven to be of additional value in many fields, especially for the detection and local staging of prostate cancer[5,6]. The three default MR sequences for prostate cancer imaging are T2-weighted MR, DW-MR, and DCE-MR, similar to the MR sequences used in this thesis for the image analysis of liver lesions. Even though the combination of these three MR sequences is not studied in this thesis, it is assumed that the exploitation of the three MR sequences has great potential for the development of more advanced techniques for the detection, segmentation, and classification of focal liver lesions.

Image registration

Registration of the MR images is required for the analysis of multiple images. The results of our attempts to align the DW-MR images with the DCE-MR images were not always of satisfying quality. Some data had to be discarded due to the unsatisfactory alignment of the images. Attempts to register T2-weighted MR images to DCE-MR images also failed. Since the liver is a very deformable organ, abdominal image registration is a challenging task. The challenges arise due to the possibility of large non-linear deformations of the liver between the two MR images, individual movement of other abdominal organs, different spacing between slices in T2-weighted images, as well as the lack of detailed anatomical information in DW-MR images. In DCE-MR images motion is limited by asking the patient to hold his/her breath several times, however it is not feasible to use this strategy in all MR sequences.

To aid the multiparametric registration, patients should be asked to breathe superficially during the acquisition of the MR sequences for which breath hold is no option. This limits the nonlinear motion of the liver as much as possible. In addition, this allows to image the liver as a whole, without slices with overlapping information or missing information between consecutive slices. Current multiparametric registration methods provide sufficient results in most data sets, but fail to align MR images when considerable motion between them is present. New multiparametric registration methods should take in account the possibility of large deformations of

the liver and other abdominal organs.

Furthermore, longitudinal scans can be used to quantify the change of liver lesions over time. Change in volume or in appearance could indicate a positive response to a treatment or the opposite, disease progression. The treatment plan could be altered according to the status of the disease. Image registration of longitudinal MR scans would aid such follow-up image analysis of lesions. However, image registration suffers from similar challenges as mentioned before, as well as from longitudinal changes in the liver's appearance and shape due to treatment. An anatomical structure within the liver that has a more consistent appearance over time, is the vascular tree. The vascular tree could guide the registration of longitudinal liver MR scans, even when some part of the liver is surgically removed. To this end, vessel segmentation is required. The centerline of these vessels can be used a landmarks or points for a point-cloud registration instead of an intensity-based registration. Such vessel-based image registration approach has been successfully applied before.[7,8]

Clinical applicability

The image processing techniques described in this thesis are developed to assist radiologists in the analysis of abdominal MR images. Even though not all developed techniques provide clinically sufficient results yet, they show potential for clinical implementation.

One of the main achievements is the clinical implementation of the groupwise registration method for abdominal contrast enhanced MR images. The original images are sent from the MR scanner to the server and the server sends the registered images to the picture archiving and communication system (PACS) for the radiologists to access. The subtraction images are also added to the PACS. Subtraction images are images for which the pre-contrast image is subtracted from the contrast enhanced images. This provides the radiologists with the opportunity to assess the contrast uptake of the abdominal tissues. For example, the subtraction images reveal the rim-enhancement of malignant tumors or other enhancement patterns which make differentiation between benign and malignant lesions easier or more certain. These subtraction images are deemed very helpful by radiologists during the examination of the DCE MR images.

The liver segmentation method described in Chapter 4 provided satisfying results and shows potential for clinical usage. The volume of the liver can easily be calculated from the liver mask and can be used for volume measurements, such as the preoperative

volumetric assessment of the liver to prevent liver failure after partial hepatectomy.[9] The segmentation method also offers easy liver volume measurements in other clinical applications. Liver volumetric measurements are often not performed due to the amount of effort in manually annotating the images. In addition, the lesion detection method could assist the user with the detection of liver metastases if there is a known primary tumor. The method highlights regions that radiologists might want to consider more closely. This could lead to a more efficient and effective radiologic assessment.

Other techniques described in this thesis still need more development. The liver lesion classification is limited by the number of different lesion types and could be improved by the inclusion of more lesions, typical and atypical, to provide the user with a broad range of possible lesions. In addition, this method should be evaluated properly with a test set with tens of patients, instead of a leave-one-patient-out principle.

Deep learning could also be applied to the classification method to learn the features of the lesions instead of handcrafting them. In addition to a sole classification, an interpretable approach is possible as presented by Wang et al. (2019)[10]. This technique does not only provide a class label, but also includes features on which the decision was based. These features can include a central scar, enhancing rim, or washout of contrast agent. This would not only give the most probable lesion class, but also shows more insight and provide a result that is better interpretable by humans. The radiologist can check the outcome of several features easily by looking at the MR images.

The described lesion detection method could be expanded to provide the user not only with a location, but with a segmentation or delineation of the lesion. The detection method could be turned into a detection and segmentation method by using post-processing methods as conditional random fields[11] or iterative corrections to the segmentation[12,13]. These methods use the detection result as an initial segmentation and improve the segmentation using local information. This local information can consist of edges, gray level intensities, or manual scribbles. If DCE-MR and DW-MR images are used for detection and segmentation, they should be registered. Proper delineation is not feasible if lesions in the two MR sequences are not aligned.

Data for deep learning

The configuration of multiple MR images as input data for a convolutional neural network (CNN) plays an important role in CNN design. Different input configurations

can have different performance levels. Accordingly, various input configurations for a CNN should be considered, provided that the data has multiple images. If MR images of one MR sequence are correlated, like the phases of the DCE-MR or DW-MR images, the CNN can extract features which combine the images in a proper way. For example, temporal features in DCE-MR images, like the time to peak or washout rate, can be extracted from the combined images. Combining the correlated images of one MR sequence with images of another MR sequence at the input of the CNN will decrease the effectivity of the extracted features, especially temporal features. It is preferred to extract the features of the two sequences before concatenation of the feature maps. This way the CNN can focus on the extraction of features from each MR sequence separately.

In Chapters 4 and 5, the same network architecture and parameters are used to train the CNN for liver metastases detection (excluding patient-specific fine-tuning). However, the two CNNs do not have a similar performance level. The main difference between the two trained CNNs is the amount of training data. The same data set is used, but the training set of Chapter 4 is partly used as the test set in Chapter 5 and vice versa. The patient-specific fine-tuning step required follow-up data, leaving less data available for training the base CNN. As a result, twice as many slices with metastases were available for training the liver metastases detection CNN in Chapter 4 than in Chapter 5. The drop in performance of the true positive rate is about 30 per cent. This is in line with the general statement that a large enough data set is required to train a CNN with satisfying results[14].

That being said, several data augmentation methods are developed to make optimal use of data for deep learning in order to reduce the amount of data required. Data augmentation manipulates the existing data in such a way that 'unique' data is created. This results in more examples for a CNN to learn from the data. In this thesis only simple data augmentation techniques are applied, i.e. rotation and flipping of image patches. More advanced data augmentation techniques involve for example elastic deformations[15], intensity inhomogeneity augmentation[16], ground truth contour augmentation[17], or synthetic medical image augmentation with generative adversarial networks[18]. The use of these techniques can boost the performance of a CNN without the addition of more data and/or annotations.

In addition to data augmentation, the order in which the data is presented to the CNN has an effect on the performance. A study by Toneva et al. (2019)[19] suggests that some examples are more easily forgotten by the network than other examples. Training first on the easy examples followed by training the CNN with the more easily

forgotten examples, or harder examples, gives a good performance for both groups of examples, rather than using them for training all at once. This way of training can also be done even more explicitly by training two consecutive CNNs[18]. For example in lesion detection, the first CNN will suggest possible lesions and labels the easy patches or pixels. The second CNN receives all the patches or pixels that are not easily distinguishable and learns to differentiate between the lesion-like patches.

Altogether, the way and order in which the data is presented to the CNN is very important to achieve good results. By exploiting the data in the right manner, better or similar results can be achieved while using less data. This should be considered when CNNs are used for image analysis tasks. The optimal usage of medical image data and the corresponding annotations for segmentation or classification tasks are a key point of investigation in medical image analysis.

Personalized medicine

In Chapter 5, we used the patient's previously acquired scans to update a lesion detection CNN towards the features of that specific patient. By personalizing the CNN, an increase in performance was achieved. This study showed the benefits of updating a pre-trained CNN with limited, but very informative data. The refinement of the CNN with this very specific data could be seen as overfitting the CNN towards that specific patient.

This approach can be beneficial in situations where a scan of a patient is acquired regularly to monitor disease or treatment progression. It moves the standard image analysis to a more personalized analysis. Especially in patients with rare diseases this might be a way to benefit from the power of deep learning, without the need of a large amount of data that represents the rare disease. With every scan added to the fine-tuning data, the CNN learns the specific features and is more certain about its decisions. This approach fits in the present idea of personalization of treatment and medication for each individual patient.

This thesis presented different image analysis methods for liver MRI and showed the power of dynamic contrast enhanced MR images in particular. The methods are a step forward to clinically useful tools in liver MR analysis.

References

1. Ben-Cohen, A. *et al.* Fully convolutional network and sparsity-based dictionary learning for liver lesion detection in CT examinations. *Neurocomputing* **275**, 1585–1594 (2018).

2. Vorontsov, E., Tang, A., Pal, C. & Kadoury, S. Liver lesion segmentation informed by joint liver segmentation. in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 1332–1335 (2018).

3. Christ, P. F. *et al.* Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. in *Lecture Notes in Computer Science MICCAI 2016* **9901**, 415–423 (2016).

4. Hu, P., Wu, F., Peng, J., Liang, P. & Kong, D. Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution. *Phys. Med. Biol.* **61**, 8676–8698 (2016).

5. Hoeks, C. M. A. *et al.* Prostate cancer: Multiparametric MR imaging for detection, localization, and staging. *Radiology* **261**, 46–66 (2011).

6. Weinreb, J. C. *et al.* PI-RADS Prostate Imaging - Reporting and Data System : 2015 , Version 2. *Eur. Urol.* **69**, 16–40 (2016).

7. Osorio, E. M. V. *et al.* Accurate CT/MR vessel-guided nonrigid registration of largely deformed livers. *Med. Phys.* **39**, 2463–2477 (2012).

8. Reinertsen, I., Lindseth, F., Unsgaard, G. & Collins, D. L. Clinical validation of vessel-based registration for correction of brain-shift. *Med. Image Anal.* **11**, 673–684 (2007).

9. Lodewick, T. M. *et al.* Fast and accurate liver volumetry prior to hepatectomy. *Int. Hepato-Pancreato-Biliary Assoc.* **18**, 764–772 (2016).

10. Wang, C. J., Hamm, C. A., Letzen, B. S. & Duncan, J. S. A probabilistic approach for interpretable deep learning in liver cancer diagnosis. in *Proc. SPIE 10949, Medical Imaging 2019: Image Processing* 109500U (2019). doi:10.1117/12.2512473

11. Christ, P. F. *et al.* Automatic Liver and Tumor Segmentation of CT and MRI Volumes Using Cascaded Fully Convolutional Neural Networks. *arXiv Prepr. arXiv1702.05970* 1–20 (2017).

12. Wang, G. *et al.* DeepIGeoS : A deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* (2018). doi:10.1109/TPAMI.2018.2840695

13. Bredell, G., Tanner, C. & Konukoglu, E. Iterative Interaction Training for Segmentation Editing Networks. in *Shi Y., Suk HI., Liu M. (eds) Machine Learning in Medical Imaging. MLMI 2018. Lecture Notes in Computer Science, vol 11046* 363–370 (Springer International Publishing, 2018). doi:10.1007/978-3-030-00919-9

14. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).

15. Castro, E., Cardoso, J. & Costa Pereira, J. Elastic deformations for data augmentation in breast cancer mass detection. in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* 230–234 (2018). doi:10.1109/BHI.2018.8333411

16. Khalili, N. *et al.* Automatic brain tissue segmentation in fetal MRI using convolutional neural networks. *Magn. Reson. Imaging* (2019). doi:10.1016/j.mri.2019.05.020

17. Javaid, U., Dasnoy, D. & Lee, J. A. Semantic segmentation of computed tomography for radiotherapy with deep learning: Compensating insufficient annotation quality using contour augmentation. in *Proc. SPIE 10949, Medical Imaging 2019: Image Processing* 109492P (2019). doi:10.1117/12.2512461

18. Frid-adar, M. *et al.* GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **321**, 321–331 (2018).

19. Toneva, M. *et al.* An empirical study of example forgetting during deep neural network learning. in *ICLR 2019 AN* 1–18 (2019).

20. Wolterink, J. M. *et al.* Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. *Med. Image Anal.* **34**, 123–136 (2016).

7

A

Appendices

Nederlandse samenvatting

Dankwoord

List of publications

Biography

Nederlandse samenvatting

De lever is het grootste interne orgaan in het lichaam en vervult vele vitale functies, waaronder de productie van eiwitten en bloedstollingsfactoren, de aanmaak van glycogeen, de productie van gal en de eliminatie van giftige stoffen. Veel ziekten hebben impact op de lever en haar functie. Verschillende ziekten doen zich voor in de vorm van focale laesies (plaatselijk aangetast weefsel) en kunnen kwaadaardig of goedaardig zijn.

Kwaadaardige laesies hebben een grote invloed op de gezondheid en de kwaliteit van leven van de patiënt. Kwaadaardige laesies zijn ofwel primaire levertumoren: hepatocellulair carcinoom (HCC), of levermetastasen. Levermetastasen zijn uitzaaiingen naar de lever vanuit een tumor ergens anders in het lichaam, voornamelijk vanuit tumoren in het maagdarmkanaal. Primaire leverkanker is wereldwijd de zesde meest voorkomende vorm van kanker en een van de meest voorkomende kankergerelateerde doodsoorzaak. Jaarlijks sterven wereldwijd ongeveer 700.000 mensen aan leverkanker. Naast deze kwaadaardige laesies zijn er ook verschillende goedaardige leverlaesies, zoals adenomen, hemangiomen, focale nodulaire hyperplasie (FNH) en cysten.

Het is zeer belangrijk dat leverlaesies op tijd worden gevonden en worden aangeduid als goedaardig of kwaadaardig, zodat de juiste behandeling kan worden gestart. Er zijn veel verschillende behandelingen mogelijk voor kwaadaardige laesies, zoals: chirurgische verwijdering, radiofrequente ablatie, transarteriële chemo-embolisatie, chemotherapie en, in ernstige gevallen, levertransplantatie. De effectiviteit van de behandeling hangt af van het type laesie, de locatie en het aantal laesies. Om deze informatie te verkrijgen worden onder andere medische beelden gemaakt door middel van echografie, computertomografie (CT) of magnetic resonance imaging (MRI). Nadat de radioloog deze beelden heeft geanalyseerd, bepaalt de medisch specialist op basis van de radiologische bevindingen welke behandeling wordt ingezet.

Hoewel CT al lange tijd de voorkeursmethode is voor het detecteren en monitoren van leverlaesies, wordt MRI steeds populairder dankzij een beter contrast tussen lever en laesie zonder het gebruik van straling. In dit proefschrift zal de nadruk liggen op de beeldanalyse van lever MR beelden.

Magnetic resonance imaging

Tijdens een lever MRI onderzoek worden verschillende MR beelden opgenomen: T1 gewogen, T2 gewogen, diffusie gewogen en een dynamische contrastversterkte serie (DCE). Deze beelden geven het lichaam allemaal op een andere manier weer en laten zo verschillende weefsels en weefseleigenschappen zien. De hele lever wordt in beeld gebracht door herhaaldelijk een dwarsdoorsnede, ook wel plak genoemd, te scannen. Samen vormen deze plakken een 3D beeld. In Figuur 1 worden voorbeelden gegeven van T1 gewogen, T2 gewogen, en diffusie gewogen MR beelden van de buik.

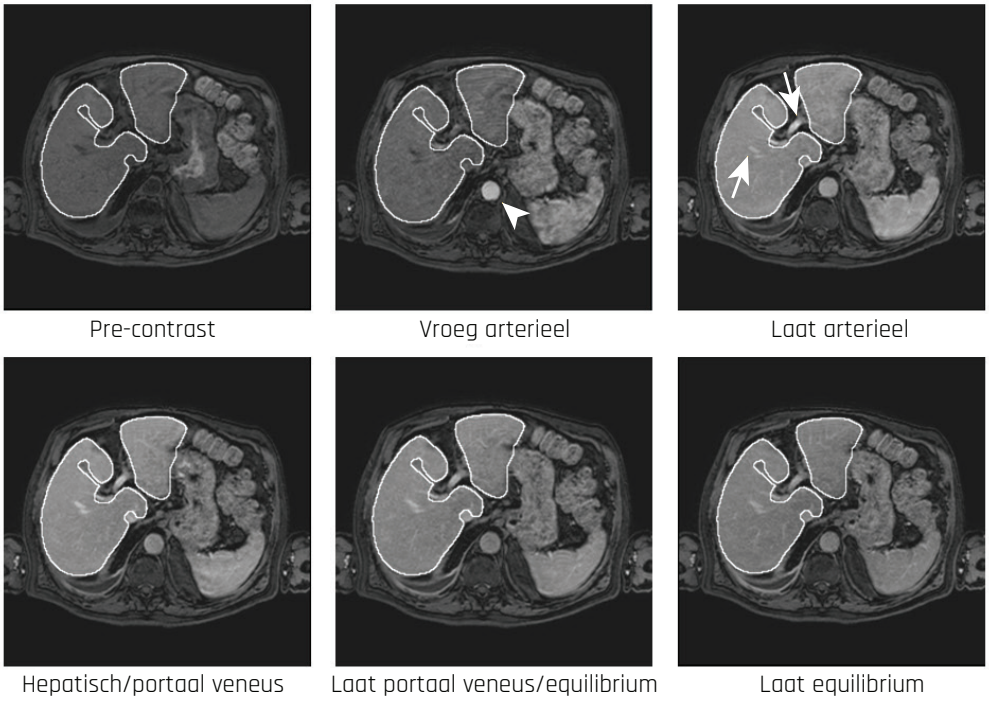
T1 gewogen MR | T2 gewogen MR | Diffusie gewogen MR

Figuur 1 Voorbeelden van T1 gewogen, T2 gewogen, en diffusie gewogen MR beelden van een gezonde lever. De lever is het grote orgaan aan de linkerkant. De galblaas licht op in de T2 gewogen en diffusie gewogen MR beelden, aangegeven met de zwarte pijl. De maag licht op in het diffusie gewogen MR beeld, aangegeven met een witte pijl. In de rechterbenedenhoek is de milt te zien (witte asterisk).

In dit proefschrift staat het gebruik van de DCE-MR beelden centraal. Deze beelden bevatten niet alleen anatomische, maar ook functionele informatie door het gebruik van een contrastmiddel. Eerst wordt een pre-contrastbeeld gemaakt, waarna het contrastmiddel intraveneus in de bloedbaan wordt toegediend. Na de toediening worden meerdere 3D beelden gemaakt om het contrastmiddel, en dus daarmee de bloedstroom, door het lichaam en specifiek door de lever te volgen. Het contrastmiddel heeft bepaalde eigenschappen om een hoge grijswaarde of intensiteit op de MR beelden te produceren, waardoor de bloedstroom een goed contrast heeft met de achtergrond. Een pixel met een hoge intensiteit is namelijk wit en een pixel met een lage intensiteit donkergrijs tot zwart.

De lever heeft een bloedtoevoer via de leverslagader en via de poortader. Hierdoor ontstaat een complex systeem, wat resulteert in verschillende contrastfasen. De fasen zijn afhankelijk van de tijd tussen de toediening van het contrastmiddel en het maken het MR beeld. De duur van de opname van de DCE-MR beelden kan oplopen tot 20 minuten. Figuur 2 toont de zes fasen van contrastversterking in MR beelden; pre-

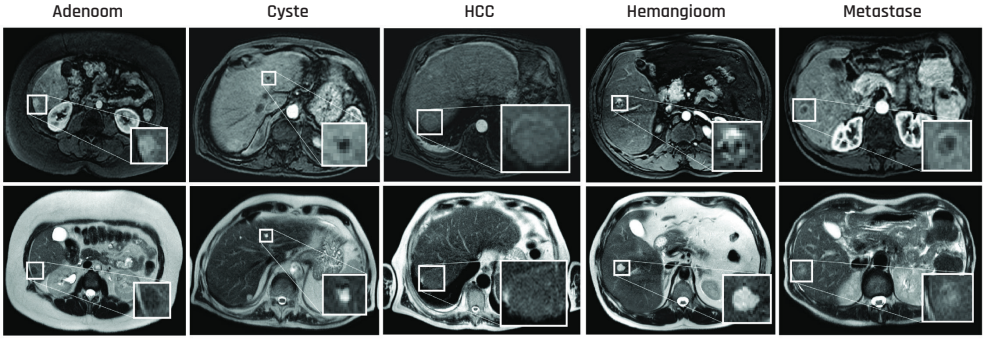
contrast, vroege arteriële fase, late arteriële fase, hepatische/portaal veneuze fase, late portaal veneuze/evenwichtsfase en de late evenwichtsfase. De MR beelden van de fasen zijn dus op verschillende tijdstippen gemaakt. Per fase worden meerdere 3D beelden gemaakt. De aankleuring van verschillende vaten en weefsels in de fasen geven een overzicht van de perfusie in de lever en weefseleigenschappen.



Figuur 2 Dwarsdoorsnede van de zes contrastfasen van de dynamische contrastversterkte MR beelden. De lever is omlijnd in wit. In het pre-contrastbeeld wordt de anatomie weergegeven zonder contrastmiddel. In de vroege arteriële fase bevat de aorta een grote hoeveelheid contrastmiddel en heeft dus een hoge intensiteit (pijlpunt). In de late arteriële fase komt het contrastmiddel via de bloedvaten (pijlen) in de lever, waardoor het leverparenchym aankleurt. Na enige tijd wordt het contrastmiddel weer uit de lever verwijderd en kleurt de lever weer donkerder.

De radioloog gebruikt de DCE-MR beelden, de andere MR beelden en patiënteninformatie om een diagnose te stellen. Alle MR beelden bevatten specifieke informatie die nuttig is om laesies te detecteren en te karakteriseren. In Figuur 3 worden vijf voorbeelden van laesies getoond in de vroeg arteriële fase van de DCE-MR beelden en het T2 gewogen MR beeld. Een hoge intensiteit op een T2 gewogen MR beeld is bijvoorbeeld een eigenschap van een cyste of hemangioom. De textuur van een laesie op een MR beeld kan ook helpen bij de karakterisering. Bijvoorbeeld de aankleuring van de rand van een HCC of levermetastase, of de heterogene aankleuring van een hemangioom op de DCE-MR serie. Beeldverwerking en

beeldanalyse kunnen de radioloog helpen bij het stellen van een diagnose.



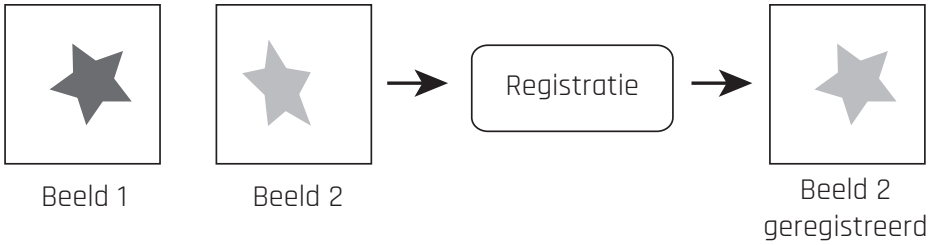
Figuur 3 Voorbeelden van leverlaesies. Van links naar rechts worden levers met een adenoom, cyste, HCC, hemangioom en een metastase van een colorectale carcinoom getoond. De bovenste rij toont de arteriële fase van de DCE-MR beelden en de onderste rij de T2-gewogen MR beelden. Een zoom-in van de laesies is toegevoegd.

Beeldregistratie

De organen in de buik en borst zijn constant in beweging: de uitzetting van de longen door ademhaling, het kloppende hart en de darmperistaltiek. Door deze natuurlijk bewegende organen wordt ook de lever continu een beetje vervormd en verplaatst. Door de ademhaling kan de lever zelfs met een paar centimeters omhoog of omlaag worden geduwd. Tijdens het MRI onderzoek worden er beelden gemaakt van verschillende plakken. Gedurende dit onderzoek is de lever in beweging. Wanneer een plak met het midden van de lever in beeld is gebracht, kan een moment later dezelfde plak de bovenkant van de lever tonen. In die korte tijd is de lever dus een aantal centimeter verplaatst. Dit zorgt ervoor dat er wazige randen ontstaan, ook wel bewegingsartefacten genoemd. Het zorgt er ook voor dat de plakken moeilijk met elkaar te vergelijken zijn, omdat ze niet dezelfde anatomie tonen. Patiënten worden daarom gevraagd om tijdens het MRI onderzoek de adem in te houden, zodat de lever op dezelfde plek blijft. Aangezien het maken van de MR beelden enige tijd duurt, voor de DCE-MR beelden zelfs tot 20 minuten, wordt de patiënt gevraagd meerdere malen de adem in te houden. Dit leidt echter niet tot beelden die helemaal vrij zijn van beweging veroorzaakt door de ademhaling. Het is namelijk niet altijd gemakkelijk voor patiënten om hun adem op dezelfde diepte of voor langere tijd in te houden.

De beweging tussen de beelden kan na het maken van de MRI gedeeltelijk worden gecorrigeerd door middel van beeldregistratie. Beeldregistratie is een methode waarbij een of meerdere beelden naar het coördinaatsysteem worden vervormd

van een referentiebeeld. Dat wil zeggen dat beelden zo worden vervormd dat de (anatomische) structuren van de beelden na registratie op dezelfde locatie liggen als die van het referentiebeeld. Zo is het voor een radioloog makkelijker om de beelden te vergelijken. Figuur 4 geeft een voorbeeld van registratie.



Figuur 4 Beeld 1 en 2 geven dezelfde ster weer, maar doordat deze beweegt is de ster in beeld 2 vanuit een andere hoek weergegeven. Beeld 1 is het referentie beeld. Door middel van registratie wordt beeld 2 zo vervormd, zodat de twee beelden meer op elkaar lijken en beter te vergelijken zijn.

In **hoofdstuk 2** zijn twee verschillende registratie methoden voor DCE-MR beelden vergeleken. De eerste methode registreert elk 3D beeld afzonderlijk naar het referentiebeeld. Als referentiebeeld is het eerste, pre-contrast, beeld van de DCE-MR beelden genomen. De tweede methode is een groepsgewijze registratie die alle beelden tegelijk naar elkaar registreert, zonder een referentiebeeld te kiezen. Deze laatste manier geeft een betere registratie zonder vreemde vervormingen en met een soepelere overgang tussen de verschillende fasebeelden. De bevredigende resultaten van de groepsgewijze registratie hebben geleid tot de implementatie van deze methode in de dagelijkse praktijk voor DCE-MR beelden van de buik. Daarnaast is de groepsgewijze registratiemethode toegepast op de DCE-MR beelden in de andere hoofdstukken beschreven in dit proefschrift.

Machine learning

De radioloog inspecteert de verschillende MR beelden, lokaliseert de laesie en bepaalt welk type laesie het is op basis van specifieke eigenschappen. Machine learning algoritmes kunnen de radioloog helpen bij het opsporen en classificeren van een laesie.

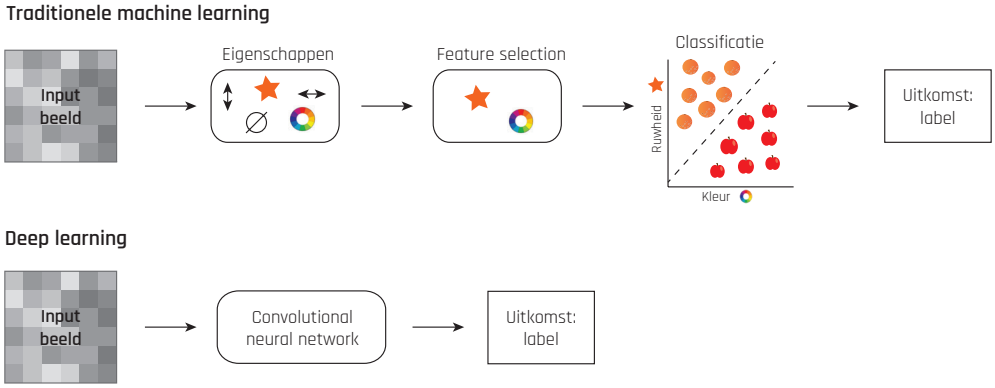
Machine learning is een verzamelnaam voor algoritmes die zichzelf leren wat de eigenschappen zijn voor een bepaald voorwerp en deze labelt. Een voorbeeld is het labelen van twee klassen: appels en sinaasappels. Twee eigenschappen die deze twee fruitklasses beschrijven zijn de kleur en de ruwheid van de schil. Een machine learning algoritme leert het verschil tussen deze twee klassen op basis van de eigenschappen.

Als het voorwerp een rode en gladde schil heeft, dan is het waarschijnlijk een appel. Als het een oranje ruwe schil heeft, is het waarschijnlijk een sinaasappel. Het algoritme leert zelf welke eigenschappen overeenkomen met welke klasse. Het leren onderscheid te maken worden ook wel trainen genoemd. Het toeschrijven van een label aan een voorwerp heet classificatie.

Het algoritme leert onderscheid te maken tussen twee of meer klassen door veel voorbeelden met het bijbehorende label te zien. Een traditioneel machine learning algoritme krijgt een lijst met eigenschappen en het klasselabel per voorwerp getoond. Deze eigenschappen zijn vaak eigenschappen die mensen ook gebruiken om onderscheid te maken tussen verschillende voorwerpen. De lijst met eigenschappen is dan ook samengesteld door een persoon.

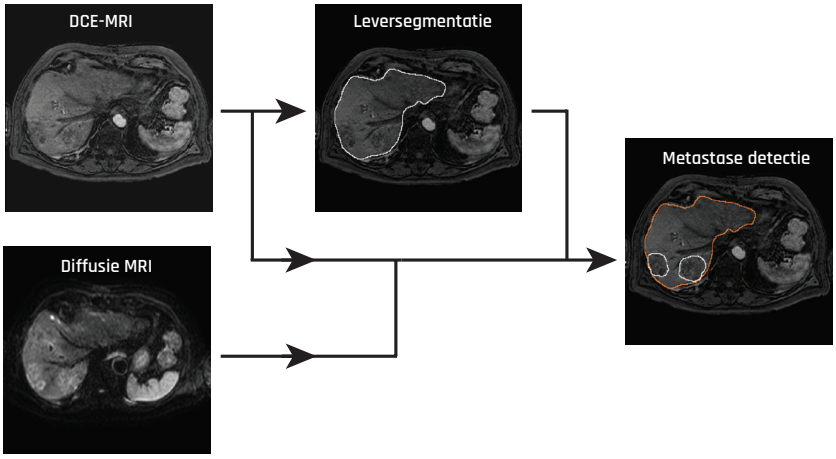
Niet alle eigenschappen zijn even belangrijk voor een machine learning algoritme om het onderscheid te maken. Om het onderscheid tussen appels en sinaasappels te maken, is de kleur bijvoorbeeld een goede indicator. Een sinaasappel is altijd oranje en een appel kan rood, geel, of groen zijn, maar is nooit helemaal oranje. De diameter van het fruit is een minder goede eigenschap om te gebruiken, omdat een appel en sinaasappel even groot kunnen zijn. Vaak worden er vele eigenschappen uitgerekend, maar zijn er eigenlijk maar een aantal van belang voor het algoritme. Om het aantal eigenschappen te beperken worden alleen de meest informatieve eigenschappen uitgekozen om te gebruiken in het leerproces. Dit wordt feature selection (eigenschapselectie) genoemd. De geselecteerde eigenschappen worden daarna gebruikt om het algoritme te trainen.

Recent zijn er machine learning algoritmes ontwikkeld die ook zelf de eigenschappen leren. Dit heet *deep learning*. Hierbij krijgt het algoritme alleen de beelden en het label te zien en niet de berekende eigenschappen. Het algoritme leert zelf wat de eigenschappen zijn. In de medische beeldanalyse worden vaak *convolutional neural networks* (CNN) gebruikt. CNN algoritmes gebruiken wiskundige convoluties om eigenschappen van beelden te berekenen en zijn een groep algoritmes binnen de deep learning groep. Deep learning algoritmes zijn succesvoller in het uitvoeren van complexe taken dan de traditionele machine learning algoritmes, omdat het zelf de eigenschappen kan berekenen. Dus ook eigenschappen die voor mensen juist niet van belang zijn, maar wel voor het algoritme. Een deep learning algoritme moet langer trainen om een goede classificatie te maken. Hiervoor zijn ook meer voorbeelden nodig dan bij de traditionele machine learning algoritmes. In Figuur 5 wordt een overzicht getoond van de twee benaderingen voor machine learning.



Figuur 5 Overzicht van een traditioneel machine learning algoritme en van een deep learning algoritme. In beide gevallen wordt een beeld in het algoritme gestopt. Bij het traditionele algoritme worden verschillende eigenschappen berekend, een selectie van die eigenschappen gemaakt en op basis daarvan leert het algoritme de twee klassen te onderscheiden. Als uitkomst krijgt een beeld een label van een van de klassen, in dit geval appels of sinaasappels. Het deep learning algoritme berekent zelf de eigenschappen en leert deze te combineren om de classificatie te maken. In medische beeldanalyse wordt vaak een convolutional neural network (CNN) gebruikt als deep learning algoritme.

In dit proefschrift zijn deep learning algoritmes gebruikt in hoofdstuk 3, 4 en 5 en een traditioneel machine learning algoritme is gebruikt in hoofdstuk 6.



Figuur 6 Overzicht van de detectiemethode. Nadat de DCE-MR beelden zijn geregistreerd, worden ze gebruikt om de lever te segmenteren. Binnen de lever vindt de detectie van levermetastasen plaats. De CNN voor detectie gebruikt hiervoor DCE-MR en diffusie gewogen MR beelden.

In **hoofdstuk 3** wordt een CNN gebruikt om alle pixels binnen de lever te labelen als lever en de pixels erbuiten als achtergrond. Dit proces heet segmentatie. In het hoofdstuk is onderzocht op welke manier de DCE-MR beelden aan het algoritme moeten worden aangeboden, om de beste leversegmentatie te verkrijgen.

In **hoofdstuk 4** is de leversegmentatiemethode verder geëvalueerd. De leversegmentatie wordt daarna gebruikt om binnen dit gebied te zoeken naar levermetastasen. Voor de detectie van deze levermetastasen wordt ook een CNN gebruikt. Als inputbeelden worden niet alleen DCE-MR beelden gebruikt, maar ook diffusie gewogen MR beelden zijn gebruikt. Door diffusie gewogen MR beelden toe te voegen aan de inputbeelden werden 50 procent meer levermetastasen gevonden. Met de combinatie van beide MR beelden als input in de CNN worden 99% van alle levermetastasen uit deze data set gevonden. Figuur 6 geeft een overzicht van de levermetastasedetectiemethode.

In **hoofdstuk 5** is de detectiemethode voor levermetastases verder ontwikkeld tot een patiëntspecifieke methode. Eerdere MRI onderzoeken van de patiënt worden gebruikt om de CNN uit hoofdstuk 4 verder te trainen en zo te verfijnen. Doordat de CNN nu ook voorbeelden ziet van die ene patiënt, wordt de CNN afgestemd op de specifieke eigenschappen van de metastasen en het omringende gezonde weefsel van die patiënt. Zo is de detectie van levermetastasen nauwkeuriger. In plaats van 67%, wordt na patiëntspecifiek trainen 85% van de levermetastasen correct gedetecteerd met minder foutpositieven. Deze getallen zijn lager dan die van hoofdstuk 4, doordat er in dit hoofdstuk minder data en dus voorbeelden aanwezig waren om de CNN te trainen. Uit dit hoofdstuk blijkt dat deze aanpak ook werkt voor andere laesies en organen, zoals voor de segmentatie van witte stof laesies in het brein.

In **hoofdstuk 6** wordt een methode beschreven voor de automatische classificatie van vijf leverlaesies. De vijf laesieklassen zijn: adenomen, cysten, hemangiomen, HCC's en metastasen. De eerste drie zijn goedaardige laesies en de laatste twee kwaadaardig. In tegenstelling tot de vorige hoofdstukken wordt er gebruik gemaakt van een traditioneel machine learning algoritme. De eigenschappen die radiologen gebruiken om onderscheid te maken tussen de leverlaesies zijn ook gebruikt om een machine learning algoritme te trainen. Bijvoorbeeld de intensiteit of de aankleuring van de laesie in een bepaalde contrastfase, zoals eerder omschreven. Feature selectie is voorafgegaan aan de training van het algoritme. Met dit algoritme kunnen de laesies in 77% van de gevallen correct worden geclassificeerd. Voor het onderscheiden van goed- en kwaadaardige laesies is dit zelfs 90%. Deze methode zal echter nog verder moeten worden ontwikkeld om betrouwbaar gebruikt te kunnen worden.

De methodes omschreven in dit proefschrift zijn ontwikkeld om radiologen te ondersteunen in de beeldanalyse van lever MRI.

Dankwoord

Graag wil ik nog een aantal mensen bedanken voor hun bijdrage aan dit proefschrift. Zonder de steun en het vertrouwen van velen was dit boekje niet tot stand gekomen.

Allereerst wil ik mijn promotoren Josien Pluim en Max Viergever bedanken. Josien, bedankt voor de vele wetenschappelijke inzichten die je me hebt gegeven, het vertrouwen, en je deskundige en kritische blik. Ondanks je drukke agenda, had je elke week tijd voor me en me de begeleiding en motivatie gegeven die ik nodig had om dit proefschrift te voltooien. Max, bedankt voor je heldere commentaar, het creeëren van een goede onderzoeksomgeving bij het ISI, en de kans om bij het ISI te promoveren.

Hugo, jij bent mijn allerbeste co-promotor! Je hebt me tijdens mijn afstudeerstage overtuigd hoe leuk de wetenschap kan zijn en nadat ik was begonnen aan mijn PhD, werd je aan mijn promotieteam toegevoegd. Je hebt me met je enthousiasme en positieve instelling de mooie kanten van een promotieonderzoek laten zien. Bedankt voor alle oxford-komma's, mevislabtrucjes, en natuurlijk de gezelligheid.

Dear Robin Strand, thank you for the opportunity to visit your group at the Centre for Image Analysis, at the Uppsala University. From the first moment, I felt very welcome and I had a great time in Sweden.

Ik wil graag alle co-auteurs bedanken. Wouter Veldhuis, Frank Wessels, Maarten Niekel, en Maarten van Leeuwen, bedankt voor jullie radiologische kennis en hulp bij het opzetten van de onderzoeken. Ashis Kumar Dhara, thank you for our interesting discussions regarding deep learning. Nick Weaver, bedankt voor je hulp en enthousiasme voor ons project. Geert Jan Biessels, bedankt voor je klinische blik en commentaar.

Ik wil alle leden van de leescommissie, prof. dr. P.A. de Jong, prof. dr. D.W.J. Klomp, prof. dr. ir. A.C. Telea, prof. dr. ir. C.H. Slump en prof. dr. ir. M. Breeuwer, bedanken voor het beoordelen van dit proefschrift.

Ook wil ik graag mijn paranimfen Julia en Marloes bedanken. Julia, bedankt voor alle mooie en hilarische momenten die we samen de afgelopen jaren hebben gedeeld. Onze week samen in Delft zal ik niet vergeten.;) Marloes, bedankt voor de koffiespauzes en je enthousiasme om me mee te nemen naar insanity of andere extreme sporten. Het

was fijn om samen de laatste maanden van onze PhDs te overwinnen.

Dear colleagues at ISI, thank you all for the wonderful time I had during my PhD! Thanks for all the coffee breaks, great times at conferences, amazing fietsenrallies, the legendary Ardenneweekend with stresspong, fabulous Christmas parties, toprope sessions, work related holidays, the best birthday ever, and many drinks and dinners!

My office roommates; Samuel, Kim, and Fenghua, thank you for the scientific discussions, support, and best of all, the friendly chats. It was a pleasure sharing an office with you guys!

Britt, Hui Shan, Mike, en Sandra, het was een eer om samen met jullie MISP 2017 te organiseren.

Maria, Marjan, Jacqueline, en Renée, bedankt voor de gezellige praatjes en natuurlijk voor het draaiende houden van het ISI. Chris, Edwin, en Gerard, bedankt voor de technische support, zonder jullie was ik niet zover gekomen.

Dear colleagues in Uppsala, thank you for making my time in Sweden unforgettable! I enjoyed being part of your group and experience mushroom picking, eating surströmming, taking the boat to Helsinki, having fika, and many more wonderful things.

Ik wil ook mijn vrienden bedanken voor de nodige afleiding. Het is fijn jullie als vrienden te hebben en ik hoop dat er nog vele leuke feestjes, gezellige etentjes, dividinsdagen, fanatieke spelletjesavonden, weekendjes weg, en (ski)vakanties mogen volgen. Simone, bedankt dat je zo'n goede vriendin bent. We hebben samen altijd heel veel lol en het is een eer jouw getuige te mogen zijn. Marloes, Kim, Carrie, Marije, en Annemiek, bedankt voor jullie vriendschap, de vele wijntjes, fijne gesprekken, en dat jullie er voor me zijn.

En natuurlijk wil ik ook mijn lieve (schoon)familie bedanken voor jullie steun en interesse. Het is fijn om zulke gezellige en gekke mensen om me heen te hebben. Een speciaal woord van dank aan mijn lieve peetoom, ome Frank, aan wie ik dit proefschrift heb opgedragen. Je hebt een zware periode achter de rug gehad waarin je vocht tegen jouw leveraandoening. In deze periode heb je ook vaak in het UMC Utrecht gelegen en ik had de mogelijkheid je dan regelmatig op te komen zoeken. Je liet me zien wat een leverziekte met je doet en daardoor realiseerde ik me goed waarvoor ik dit onderzoek heb gegaan. Ondanks de reden waarom je in het UMC

Utrecht lag, heb ik genoten van deze bezoekjes vol optimisme, en natuurlijk jouw grapjes. En je optimisme was terecht, want gelukkig heb je een paar maanden geleden een nieuwe lever gekregen. Ook daar ben ik dankbaar voor en nu kunnen we genieten van Frank 2.0.

Lieve Monique, Carlijn en Jorieke, bedankt voor alle leuke zussenuitjes, met vaak een wijntje erbij. We kunnen samen altijd goed lachen en jullie houden me met beide benen op de grond. Ik had me geen betere zusjes kunnen wensen.

Lieve papa en mama, jullie staan altijd voor me klaar en ik kom altijd met veel plezier terug naar de boerderij in het mooie Brabant. Jullie hebben me altijd aangemoedigd mijn best te doen en de dingen waaraan ik begon af te maken. Dat is gelukt! Bedankt voor jullie aandacht, interesse, en liefde.

Lieve Erwin, mijn allerbeste maatje, jij betekent heel veel voor me! Een knuffel en een kus van jou maakt mijn dag altijd weer goed. Ik bewonder je eeuwige vrolijkheid en het enthousiasme voor alles waar je aan begint. Bedankt voor je liefde en dat je er altijd voor me bent. Ik kijk uit naar onze toekomst samen.

List of publications

Publications in international journals

M.J.A. Jansen, H.J. Kuijf, A.K. Dhara, N.A. Weaver, G.J Biessels, R. Strand, and J.P.W. Pluim, "Patient-specific fine-tuning of CNNs for follow-up lesion quantification" *Submitted*

M.J.A. Jansen, H.J. Kuijf, M. Niekel, W.B. Veldhuis, F.J. Wessels, M.A. Viergever, and J.P.W. Pluim, (2019) "Liver segmentation and metastases detection in MR images using convolutional neural networks", *J Medl Imag.* 6(4), 044003. doi: 10.1117/1.JMI.6.4.044003

M.J.A. Jansen, H.J. Kuijf, W.B. Veldhuis, F.J. Wessels, M.A. Viergever, and J.P.W. Pluim. (2019) Automatic classification of focal liver lesions based on MRI and risk factors. *PLOS ONE* 14(5): e0217053. doi: 10.1371/journal.pone.0217053

M.J.A. Jansen, H.J. Kuijf, W.B. Veldhuis, F.J. Wessels, M.S. van Leeuwen, and J.P.W. Pluim. (2017) Evaluation of motion correction for clinical dynamic contrast enhanced MRI of the liver. *Phys Med Biol* 62(19):7556–68. doi: 10.1088/1361-6560/aa8848

Publications in international conference proceedings

M.J.A. Jansen, H.J. Kuijf, and J.P.W. Pluim. Optimal input configuration of dynamic contrast enhanced MRI in convolutional neural networks for liver segmentation. In: *Proc SPIE 10949, Medical Imaging 2019: Image Processing*. 2019. p. 1094941V.

M.J.A. Jansen, H.J. Kuijf, and J.P.W. Pluim. Automatic classification of focal liver lesions based on clinical DCE-MR and T2-weighted images: A feasibility study. In: *IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018*. 2018. p. 245–8.

M.J.A. Jansen, W.B. Veldhuis, M.S. van Leeuwen, and J.P.W. Pluim. Motion correction of dynamic contrast enhanced MRI of the liver. In: *Proc SPIE 10133, Medical Imaging 2017: Image Processing.* 2017 p. 101331T

Biography

Mariëlle Jansen (22 October 1991, Oss) obtained her BSc in Technical Medicine cum laude in 2013 and her MSc in Biomedical Engineering in 2015 from the University of Twente. During her studies, she gained research experience during an internship at the Medical Ultrasound Imaging Centre, Radboud University Medical Centre Nijmegen. The research for her master thesis was conducted at the Image Sciences Institute, University Medical Center Utrecht.

Mariëlle started her PhD project at the Image Sciences Institute in 2015. During her PhD project she was a guest PhD student at the Center for Image Analysis, Uppsala University, Sweden. The results of her PhD project are presented in this thesis.