UNIVERSITÄT ZU LÜBECK
INSTITUTE OF MATHEMATICS AND
IMAGE COMPUTING

Development of a Vessel Segmentation Model via Deep Learning for the Analysis of Liver Function

Entwicklung eines Modells zur Gefäßsegmentierung mittels Deep Learning zur Analyse der Leberfunktion

Masterarbeit

im Rahmen des Studiengangs
Mathematik in Medizin und Lebenswissenschaften
der Universität zu Lübeck

Vorgelegt von
Rebekka Fehling

Ausgegeben und betreut von
Prof. Dr. rer. nat. Jan Lellmann
Institute of Mathematics and Image Computing

Mit Unterstützung von
Prof. Dr.-Ing. Andrea Schenk
Fraunhofer MEVIS

11. Juli 2025

MIC

IM FOCUS DAS LEBEN

Eidesstattliche Erklärung

Ich versichere an Eides statt, die vorliegende Arbeit selbstständig und nur unter Benutzung der angegebenen Quellen und Hilfsmittel angefertigt zu haben.

Lübeck,
11. Juli 2025

R.Fehling

Rebekka Fehling

Abstract

Vessel segmentation is an important tool in medical applications such as diagnosis, treatment delivery and prognosis formulation and evaluation. It is also a central part of the research project presented in this work, which focuses on assessing liver function in patients with primary sclerosing cholangitis based on magnetic resonance data of patients. The aim is to investigate whether an automated approach of liver function estimation can replace a manual region-of-interest-based analysis. In this context, vessel segmentation helps to obtain a segmentation of the pure liver parenchyma, leading to a more accurate computation of local liver function. This thesis presents the development of an algorithm for automatic liver vessel segmentation which can be applied within this project. Furthermore, the developed model as well as the whole workflow for liver function computation is evaluated in comparison to manual measurements. The results of the automatic approach show high correlation with the manual measurements, indicating that this is a promising approach for the analysis of hepatobiliary function.

Kurzfassung

Die Segmentierung von Blutgefäßen ist ein wichtiger Bestandteil in vielen medizinischen Anwendungen, wie Diagnose, Beurteilung des Behandlungsverlaufs und bei der Erstellung und Auswertung von Prognosen. Sie ist auch ein zentraler Teil des Forschungsprojektes, das in dieser Arbeit präsentiert wird und sich mit der Bewertung der Leberfunktion anhand von Magnetresonanz-Tomographie Daten von Patienten mit primär sklerosierender Cholangitis beschäftigt. Ziel ist es, herauszufinden, ob ein automatisierter Ansatz zur Schätzung der Leberfunktion die manuelle "Region-of-Interest"-basierte Analyse ersetzen kann. In diesem Kontext kann mithilfe einer Gefäßsegmentierung das reine Lebergewebe segmentiert werden, was eine genauere Berechnung der lokalen Leberfunktion ermöglicht. Diese Arbeit beschreibt die Entwicklung eines Algorithmus zur automatisierten Segmentierung der Leber-Gefäße für die Anwendung in diesem Projekt. Außerdem wird das entwickelte Modell sowie der gesamte Workflow zur Leberfunktionsberechnung im Vergleich zu manuellen Messungen ausgewertet. Die Ergebnisse des automatisierten Ansatzes liefern hohe Korrelationswerte mit den manuellen Messungen. Dies zeigt, wie vielversprechend dieser Ansatz für die Analyse der Leberfunktion ist.

Acknowledgements

This master thesis was created as part of participation in the graduate program DigiStrucMed 2020_EKPK.20, which is supported by the Else Kröner-Fresenius-Stiftung.

I would like to thank my supervisors Jan Lellmann and Andrea Schenk as well as the entire Fraunhofer MEVIS Team with Felix Thielke, Eike Petersen and Sonja Wichelmann for their outstanding support throughout this project.
Furthermore, I extend my thanks to Sina Dornbusch and Kristina Ringe from the Hannover Medical School for the great interdisciplinary collaboration.

Contents

Chapter 1: Introduction

1.1 Motivation

The assessment of liver function plays an important role in diagnosing patients with chronic liver disease such as *primary sclerosing cholangitis* (PSC). In recent years, the computation of liver function based on manually drawn *regions of interest* (ROIs) in *magnetic resonance imaging* (MRI) data has emerged as an alternative to medical scores based on blood values. This is a promising approach, as it provides not only global but also local information of the liver function as visualized in Figure 1. In this project, it is investigated whether an automatic workflow for the computation of liver function on MRI data can replace the ROI-based analysis. This thesis is complemented by the work of a medical student who performed the manual measurements of liver function and the evaluation of the automated workflow from a clinical point of view.

In the context of the automatic workflow, vessel segmentation plays an important role. The main focus of this Master's thesis is the development of a vessel segmentation model based on *Deep Learning* (DL), optimizing the automatic workflow for liver function computation. The output of the model is aimed to be subtracted from a liver segmentation mask to provide a segmentation of pure liver tissue. This application differs from the aim of existing vessel segmentation models that are mainly developed to visualize the course of the vessel structure, e.g., for operation planning. The new application requires to prefer overestimation of the segmentation to underestimation. This is the main challenge addressed in the development of the new model. Other challenges addressed in this thesis are the complex structure of the vessels and the application of the model on MRI data, in contrast to other acquisition processes, as this comes with intensity inhomogeneities that render learning more difficult.

1.2 Contribution

The following section 1.3 provides a brief overview of the medical background to this project. In particular, the function of the liver as a human organ and important aspects of its structure relevant for the following chapters are addressed. Furthermore, some information on the liver disease PSC is given. In section 1.4, the current state of research for the assessment of liver function in PSC patients is summarized. Section 1.5 presents the MRI data of PSC patients and control group used for the evaluation of the liver function computation and the dataset used for the training and evaluation of the vessel segmentation models. In section 1.6, the computational environment is described including the different steps in the CuraMate workflow for automatic liver function computation.

The main part of this thesis is the development of a vessel segmentation model adapted to the application in the workflow for computation of liver function. In chapter 2, the development process is explained in detail. In section 2.1, the motivation and challenges of the segmentation of vessels for the given application are presented. In section 2.2, an overview of existing vessel segmentation models is given. The following two sections

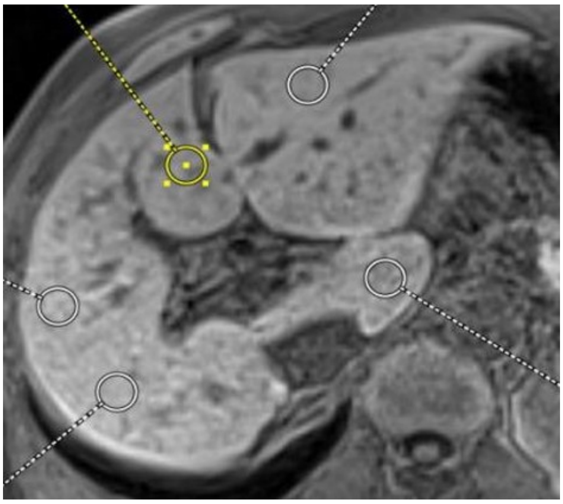


Figure 1: On MR images not only global but also local liver function can be assessed via manual set ROIs (image provided by Sina Dornbusch).

contain detailed information about the methods used for the development of the segmentation model. For this, a baseline threshold segmentation is investigated in section 2.3. As another more promising approach, a U-Net is trained with different preprocessing methods and parameters as explained in section 2.4. Section 2.5 concerns the evaluation of the vessel models on test data and the selection of the best models for additional evaluation by the medical student. Based on this evaluation, the models are advanced with two improvement methods described in a subsection of section 2.5.

In chapter 3, the best vessel segmentation models are integrated into the full workflow, and their impact on the computation of liver function is evaluated. Furthermore, the correlation results of the automatic approach with the manual measurements are presented.

The last chapter 4 summarizes the evaluation and results of this work and provides an outlook on potential future improvements.

1.3 Medical Background

1.3.1 Function and Structure of the Liver

The human liver fulfills many important functions including metabolism, production of bile, filtration and storage of blood, storage of nutrients and blood clotting [14]. The anatomical structure of the liver plays a role in later sections of this work, particularly with regard to its vascular system and the division of the liver into segments. Therefore, a short overview is given in this section.

Three blood systems run through the liver: The *portal vein* (pv) carries blood from gut, spleen and pancreas to the liver, the *hepatic artery* runs parallel to the pv carrying oxygenated blood to the liver, and the *hepatic vein* (hv) drains blood from the liver into

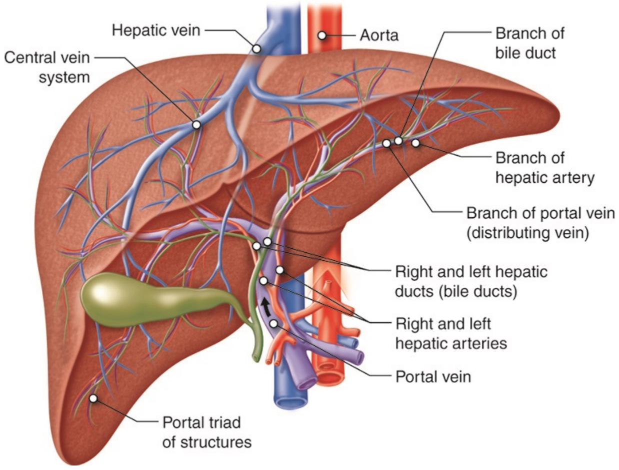

Hepatic vein
Aorta
Branch of bile duct
Central vein system
Branch of hepatic artery
Branch of portal vein (distributing vein)
Right and left hepatic ducts (bile ducts)
Right and left hepatic arteries
Portal vein
Portal triad of structures

Figure 2: The liver contains three blood systems – the portal vein, the hepatic artery, and the hepatic vein (image source: *"Cenveo - Drawing Liver anatomy and vascularisation - English labels" at AnatomyTOOL.org by Cenveo, license: Creative Commons Attribution*).

the vena cava. The *bile ducts* (bd) run parallel to the portal vein as well, which plays a role for the manual annotations described in section 1.5.2 [14]. Figure 2 visualizes the vessel structure.

In the context of this work, not only global but also local liver function is considered, for which the liver is divided into smaller segments. Couinaud classifies the liver into eight segments based on the vascular structure of the portal and hepatic vein systems [8]. This classification is visualized in Figure 3.

The eight territories according to Couinaud are combined to four segments or two liver lobes in [55]. These classifications are presented in Table 1. It should be noted that liver segment I is not included in the division into four sections, but is considered separately.

Anatomical term	**Couinaud segments**
Left lateral section	II, III
Left medial section	IVa, IVb
Right anterior section	V, VIII
Right posterior section	VI, VII
Left liver	I – IV
Right liver	V – VIII

Table 1: Relation between the eight liver segments according to Couinaud and the division into four liver sections or two liver lobes (see also Figure 3).

In this project, all three described classifications are applied and evaluated to balance segmentation quality and accuracy of the local assessment of liver function. The motivation and evaluation are presented in more detail in chapter 3.

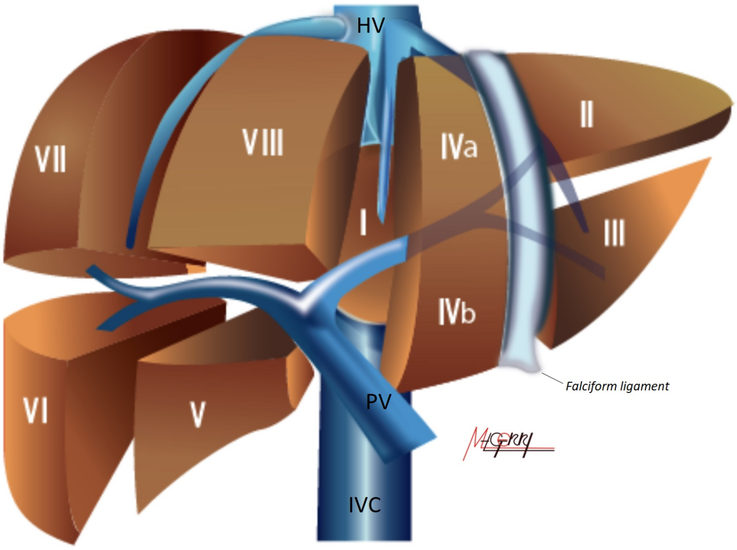


Figure 3: The liver can be divided into eight segments by Couinaud based on the vessel structure of the portal vein (PV) and the hepatic vein (HV). IVC denotes the inferior vena cava (image source: *"Radiopaedia - Drawing Hepatic segments - Number labels" at AnatomyTOOL.org by Azza Elgendy, license: Creative Commons Attribution-NonCommercial-ShareAlike;* labeling of the vessel systems added for this thesis).

1.3.2 Primary Sclerosing Cholangitis

PSC is a chronic progressive liver disease that comes with inflammation of the bile ducts. The illness is poorly understood so far, the causes are still unknown and the variable course of the disease makes it difficult to diagnose. It is therefore also known as "black box of hepatology" [37]. No medical therapy has been found for PSC yet, there is only liver transplantation as last resort in case of malignancy [37]. For diagnosis, MRI offers the possibility to assess changes in the bile ducts in patients with PSC and to visualize changes in the liver parenchyma, making it an important diagnostic tool for PSC [3].

1.4 Related Work: Liver Function

In clinical practice, the liver function of PSC patients at an advanced stage of the disease can be assessed by medical scores [37]. In recent years, the *model for end-stage liver disease* (MELD) and the Mayo score have been widely used. These are based on various blood parameters such as bilirubin, albumin or serum creatinine [23, 25]. These medical scores can especially help to determine the stage of the disease and the timing of a liver transplantation [37].

In recent years, contrast-enhanced MRI has been investigated as an alternative method for determining liver function. MRI is chosen as acquisition method since it provides a good visualization of organs and soft tissues. In contrast to the medical scores mentioned above, MRI allows assessing not only global but also local liver function. This is a great advantage, as in PSC patients, certain liver segments can be more affected than others [3].

In [2], different methods for signal intensity-based liver function measurement are proposed, including *Relative Enhancement* (RE) and Normalization to a reference organ such as the spleen or a muscle. They are based on the signal intensities of the MRI volumes in native phase (SI$_{pre}$) and of the volumes taken after contrast agent administration (SI$_{post}$). In this thesis, two liver function scores are of particular relevance. Firstly, the change in signal intensity:

$$SI_{post} - SI_{pre}. \tag{1}$$

Secondly, the relative enhancement (RE), which is defined in [2] as

$$\frac{SI_{post} - SI_{pre}}{SI_{pre}}. \tag{2}$$

For measuring the signal intensities, Schulze et al. [48] describes the method of drawing small ROIs manually in tissue-only regions of the liver. These ROIs can be defined in all liver segments (as classified by Couinaud, see section 1.3.1) to assess local liver function. Such measurements are made and evaluated as part of this project by the medical student (see Figure 1) and compared to the measurements presented in [48] to evaluate inter-reader variability and the value of the results as long-term prognostic biomarker in patients with PSC. The manual measurements show good correlation with several blood parameters and clinical outcomes. These results will be published in the medical part of this project. In this thesis, it is considered if this manual approach can be replaced by an automatic workflow. For this, the manual measurements serve as ground truth. The process of determining manual measurements is visualized in the diagram in Figure 4.

1.5 Data and Manual Ground Truth

1.5.1 Data for Liver Function Estimation

For the evaluation of liver function estimation, a historical patient cohort (March 2012 until March 2016) with 111 patients diagnosed with PSC (83 male, 28 female; mean age 45 years) has been provided by the Hannover Medical School [48]. For each patient, the dataset includes two T1-weighted MRI volumes, a native sequence before and one taken approximately 20 minutes after administration of the contrast agent (gadoxetate disodium). Additionally, 141 patients (93 female, 48 male; mean age 47 years) were chosen who fulfill the following necessary criteria to qualify as control group (February 2012 until January 2025): These are liver and kidney healthy patients at the age of 18 or older where both MRI phases have been taken and the contrast agent gadoxetate disodium has been used.

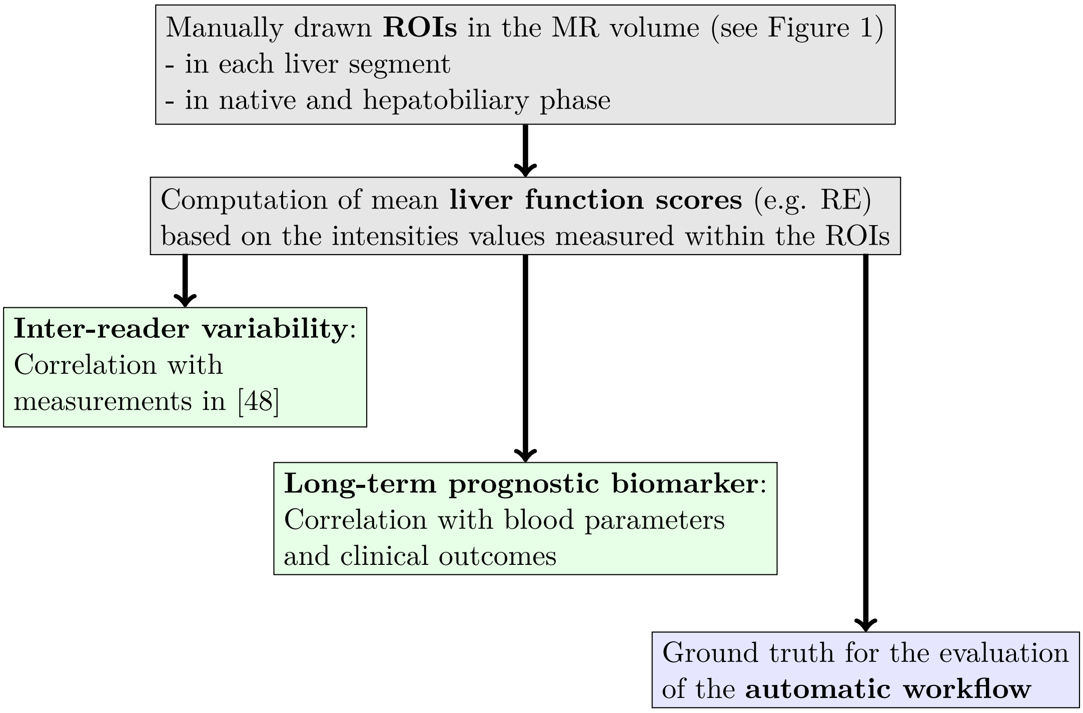

Manually drawn **ROIs** in the MR volume (see Figure 1)
- in each liver segment
- in native and hepatobiliary phase
Computation of mean **liver function scores** (e.g. RE) based on the intensities values measured within the ROIs
Inter-reader variability: Correlation with measurements in [48]
Long-term prognostic biomarker: Correlation with blood parameters and clinical outcomes
Ground truth for the evaluation of the **automatic workflow**

Figure 4: Process of determining manual measurements of liver function on MR images. The aspects in the green boxes are evaluated by the medical student of this project. The relation of the manual measurements to the automatic workflow (blue) is evaluated in this thesis.

On both datasets, signal intensity measurements were performed manually by drawing ROIs in every liver segment at corresponding positions before and after contrast agent administration. For PSC data, two independent measurements by two readers were performed. Furthermore, data on associated blood values, MELD score results and clinical outcome is included.

For the evaluation of the automatic estimation of liver function in this work, the manual measurements are used as ground truth values. The correlation with several PSC specific surrogate parameters and clinical endpoints has been shown on PSC data in the medical part of this project (not published yet).

1.5.2 Data for Training of the Vessel Segmentation Models

The SIRTOP dataset that is used in the supervised learning algorithm for the development of a vessel segmentation model was originally provided by the Municipal Hospital Dresden (internal dataset, utilized in [29]). It contains 69 MRI volumes and 93 *computer tomography* (CT) volumes in the hepatobiliary phase (after contrast agent administration) obtained for patients scheduled to undergo a *Selective Internal Radiation Therapy* (SIRT). These data contain various cases with large lesions, as SIRT is mainly performed on non-operable tumors [29]. This supports learning of image structures on data of patients with liver disease. The dataset is annotated with masks for the two largest vascular systems of the liver – hepatic and portal vein – and includes the bile ducts and hepatic artery running parallel to the portal vein.

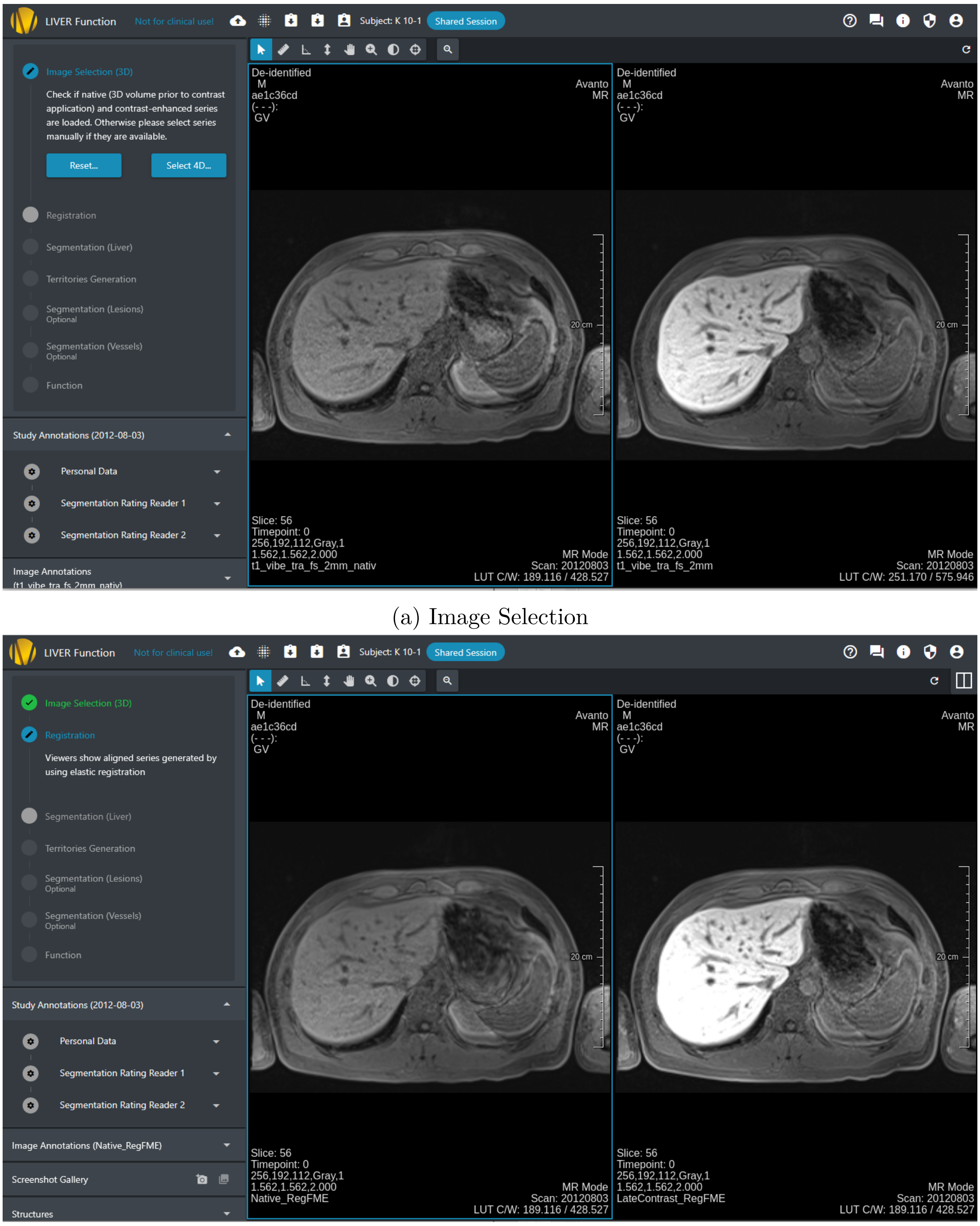
1.6 Computational Environment

For the computation of liver function, a team of Fraunhofer MEVIS implemented the basic workflow within the browser-based toolkit CuraMate (formerly SATORI) [28]. In this project, it has been further developed and now consists of the following steps: First, the two MRI volumes of native phase (before contrast agent administration) and the hepatobiliary phase (about 20 minutes after contrast agent administration) of the patient dataset are selected and loaded into the viewer. Then, the native sequence is registered to the hepatobiliary phase with elastic registration [56, 54].

For the segmentation of the liver, the DL model proposed by Haensch et al. in [15] is applied. The calculation of liver territories is done using a DL algorithm which is trained on annotated data, taking the liver segmentation as input. It is contour-based and depends on the quality of the liver segmentation. The computation of the four liver sections and the liver lobes are based on the same algorithm and merge the corresponding segments (as described in section 1.3.1).

In the next step, lesions and vessels are segmented. For the segmentation of lesions, there is a semi-automatic One-Click segmentation algorithm that computes the segmentation based on a manually set seed point [16, 17]. The vessel segmentation model has been developed in this thesis and is described in more detail in chapter 2.

The last step is the computation of liver function values based on the lesion- and vessel-free segmentation of the liver. This is evaluated for each liver segment with the two formulas for subtraction of signal intensities and relative enhancement (defined in 1.4). Figure 5 shows snapshots of the workflow visualizing the different steps.



Check if native (3D volume prior to contrast application) and contrast-enhanced series are loaded. Otherwise please select series manually if they are available.

De-identified
M
ae1c36cd
(- - -):
GV

Avanto
MR

Slice: 56
Timepoint: 0
256,192,112,Gray,1
1.562,1.562,2.000
t1_vibe_tra_fs_2mm_nativ

MR Mode
Scan: 20120803
LUT C/W: 189.116 / 428.527

Slice: 56
Timepoint: 0
256,192,112,Gray,1
1.562,1.562,2.000
t1_vibe_tra_fs_2mm

MR Mode
Scan: 20120803
LUT C/W: 251.170 / 575.946

(a) Image Selection

Viewers show aligned series generated by using elastic registration

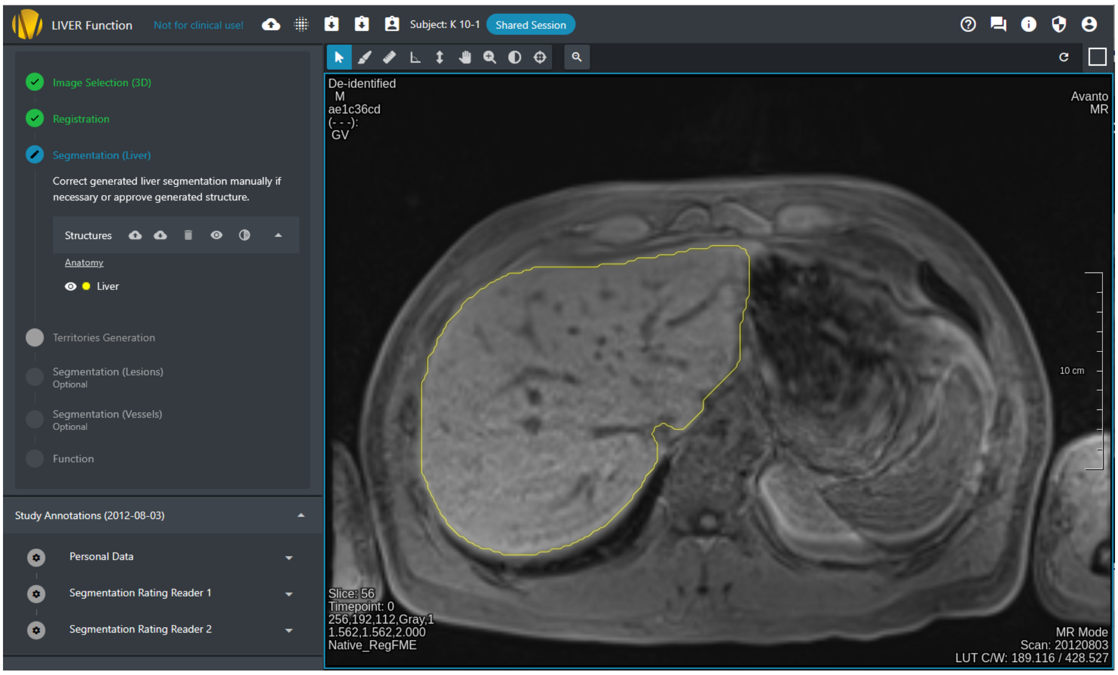
Slice: 56
Timepoint: 0
256,192,112,Gray,1
1.562,1.562,2.000
Native_RegFME

MR Mode
Scan: 20120803
LUT C/W: 189.116 / 428.527

Slice: 56
Timepoint: 0
256,192,112,Gray,1
1.562,1.562,2.000
LateContrast_RegFME

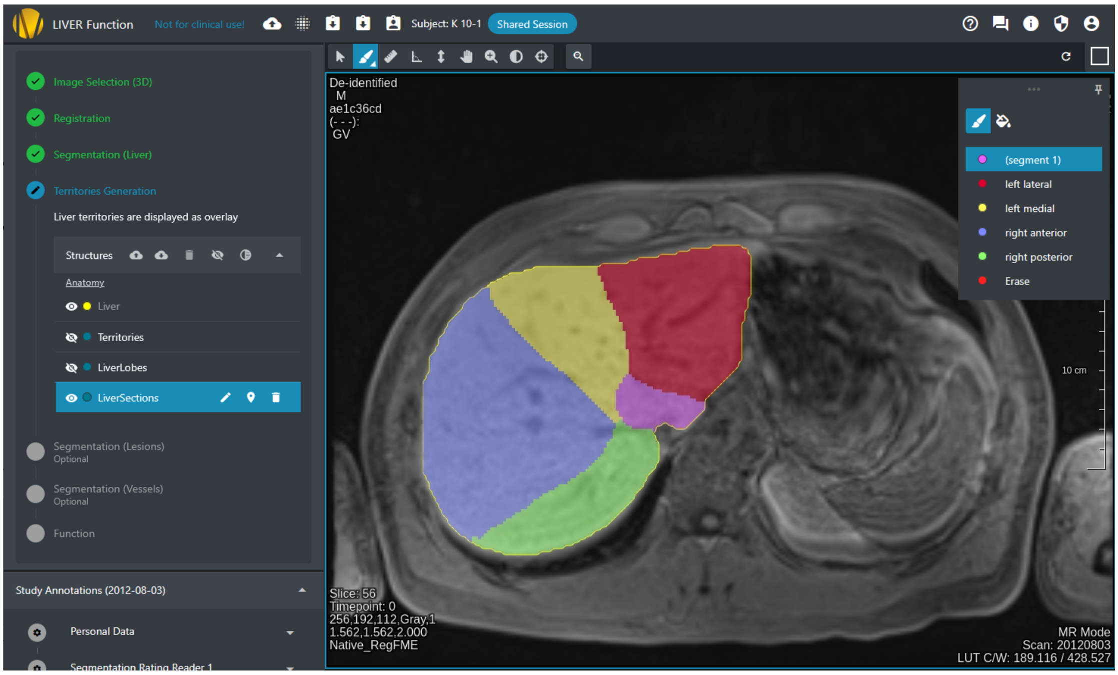
MR Mode
Scan: 20120803
LUT C/W: 189.116 / 428.527

(b) Registration



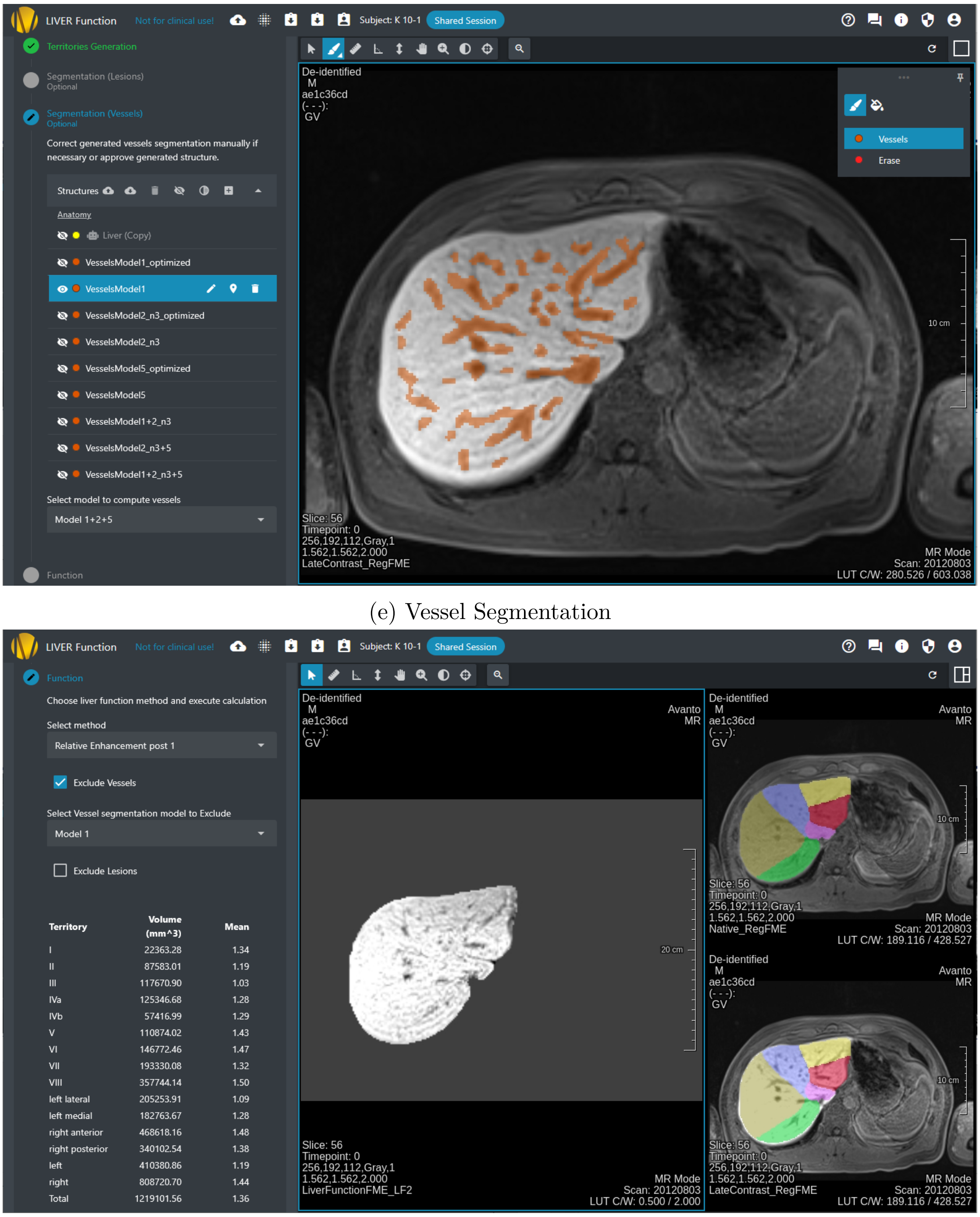
Correct generated liver segmentation manually if necessary or approve generated structure.

(c) Liver Segmentation



Liver territories are displayed as overlay

(d) Territories Calculation



LIVER Function Not for clinical use! Subject: K 10-1 Shared Session

✓ Territories Generation

Segmentation (Lesions)
Optional

Segmentation (Vessels)
Optional

Correct generated vessels segmentation manually if necessary or approve generated structure.

Structures

Anatomy

- Liver (Copy)
- VesselsModel1_optimized
- VesselsModel1
- VesselsModel2_n3_optimized
- VesselsModel2_n3
- VesselsModel5_optimized
- VesselsModel5
- VesselsModel1+2_n3
- VesselsModel2_n3+5
- VesselsModel1+2_n3+5

Select model to compute vessels

Model 1+2+5

Function

(e) Vessel Segmentation

LIVER Function Not for clinical use! Subject: K 10-1 Shared Session

Function

Choose liver function method and execute calculation

Select method

Relative Enhancement post 1

☑ Exclude Vessels

Select Vessel segmentation model to Exclude

Model 1

☐ Exclude Lesions

Territory	Volume (mm^3)	Mean
I	22363.28	1.34
II	87583.01	1.19
III	117670.90	1.03
IVa	125346.68	1.28
IVb	57416.99	1.29
V	110874.02	1.43
VI	146772.46	1.47
VII	193330.08	1.32
VIII	357744.14	1.50
left lateral	205253.91	1.09
left medial	182763.67	1.28
right anterior	468618.16	1.48
right posterior	340102.54	1.38
left	410380.86	1.19
right	808720.70	1.44
Total	1219101.56	1.36

(f) Computation of the Liver Function Scores

Figure 5: The CuraMate workflow contains all processing steps from selecting the native and hepatobiliary phase MRI volumes to the computation of liver function values based on the segmentations determined in the steps in between.

Chapter 2: Vessel Segmentation

2.1 Motivation

The automatic computation of liver function on pure liver tissue requires a lesion- and vessel-free segmentation of the liver. Existing liver and lesion segmentation approaches are applied as described in section 1.6. The model for automatic segmentation of liver vessels on MRI is developed in this work. Compared to CT, there are much less vessel segmentation models available for MRI so far [32, 5] and the application of vascular segmentation as a tool to calculate pure liver tissue differs from the applications presented in other research projects [5].

Three major challenges emerged during the development of the vascular model in this work:

1. The geometry of vessels is more complex than other organs. Thus, their long thin structures with different thickness and various branches are more difficult to segment accurately [13].

2. The development of a segmentation model on MRI data is more challenging than for other modalities such as CT. This is due to the acquisition process where inhomogeneities in the magnetic field can lead to varying signal intensities resulting in a nonuniform intensity distribution in MRI images [13].

3. The vessel segmentation model needs to fit to the application task as described above. In this case, over-estimation rather than under-estimation is preferred to remove as much vascular structure as possible from the tissue. This prevents the segmentation of liver tissue remaining after subtraction from containing vessels that could affect the subsequent calculation of liver function values.

This chapter explains how these challenges can be overcome with different data processing methods and well-chosen parameters when training a U-Net based model.

2.2 Related Work: Vessel Segmentation

Vessel segmentation models are used in all stages of medical practice: diagnosis, treatment planning, during the procedure, and for the evaluation of the treatment outcome. A main application field of hepatic vascular segmentation models is the treatment of liver cancer, where the knowledge of the exact vessel structure can help in planning surgical procedures or radiotherapy [5]. For example in [29], an ML model for segmentation of hepatic artery was trained on CT data for planning a selective internal radiation therapy. Other clinical application fields are the planning of liver transplantation [42] or mixed reality technologies that offer live images to surgeons while operating [53]. In all these applications, the focus is on the accuracy of the skeleton structure of the vessels to understand the topology. The aim in this study differs from those applications, as the ultimate purpose of the model output is to act as an additional mask on the liver segmentation.

Review papers [5, 40, 7] show a variety of vessel segmentation methods and differentiate between *Machine Learning* (ML) models and non-ML models. Methods without ML are mainly Hessian-based filters, threshold methods, region-growing and tracking methods. Those approaches have been increasingly outperformed by ML-based methods in recent years as comparisons between the results of the different methods have been shown [5, 24]. In [24], the authors compare two different methods for hv and pv segmentation on CT data. It is shown that their DL-based algorithm (CNN) has superior accuracy compared to the tracking-based algorithm and saves time. Both methods are based on the three steps of semantic segmentation, extraction of the center voxels and a tracking step.

Vessel segmentation models trained on CT data reach *Dice Similarity Coefficient* (DSC) scores of 0.90 for portal vein and 0.94 for hepatic vein [24]. In comparison, DL models based on MRI reach DSC scores of 0.63 ± 0.09 (pv) and 0.53 ± 0.12 (hv) in [64] where a nnU-Net was trained on single phase MRI and DSC values of 0.61 ± 0.03 (pv), 0.70 ± 0.05 (hv) and 0.58 ± 0.15 (bd) in [43] with training a Residual U-Net on contrast-enhanced MRI. A very large DSC for MRI models with 0.87 ± 0.01 is reached in [65] with a *two-stage two-stream graph attention U-Net* (TTGA U-Net). This U-Net contains a CNN that is trained to locate vessels as first-stage network. With a 3D simple linear iterative clustering algorithm, a graph structure is generated from the segmentation results. The second-stage network extracts graph node features through two parallel branches of a graph spatial attention network, weighting different image areas and a graph channel attention network, weighting different image features. These two streams are merged with a feature fusion module replacing skip connections of the U-Net. The DSC scores of all models mentioned above are listed in Table 2. It should be noted that the models were presented in different papers and were therefore trained and evaluated on different datasets.

Model	**Modality**	**DSC scores**		
U-Net [24]	CT	pv: 0.90,	hv: 0.94	
nnU-Net [64]	MR	pv: 0.63 ± 0.09,	hv: 0.53 ± 0.12	
Res. U-Net [43]	MR	pv: 0.61 ± 0.03,	hv: 0.70 ± 0.05,	bd: 0.58 ± 0.15
TTGA U-Net [65]	MR	0.87 ± 0.01		

Table 2: Comparison of DSC scores of existing vessel segmentation models. The TTGA U-Net shows the best DSC scores among the models for MR images.

In recent years, different types of U-Nets such as Residual U-Net [43], nnU-Net [19, 64] as well as other foundation models [61] have emerged. In this work, the 3D U-Net architecture proposed in [6] is chosen as a basis for considering and evaluating different preprocessing methods and loss functions.

2.3 Method: Threshold Segmentation

As basic method, in this work, a threshold segmentation on the SIRTOP dataset is tested. Typically, a threshold is applied to the subtraction between the registered native and hepatobiliary images. As there is no native phase of SIRTOP data available, the arterial phase is used to be subtracted and the liver mask is applied to create a vessel segmentation. Both phase volumes are normalized before subtraction. In Figure 6, the result is visualized for different thresholds. The segmentation masks show many voxels that are incorrectly labeled as vessels (false positives) and with increasing threshold the number of voxels that are erroneously labeled as background (false negatives) increases. Several vascular branches are discovered leading to mean metric values of, e.g., 0.13 for recall as well as for precision for a threshold of 0.55. The box plots on the right of Figure 6 are created on a dataset containing 15 patients of SIRTOP data which is also used as test dataset in section 2.4. The choice of the metrics is explained in more detail in section 2.5.1. Clearly, there is potential for improvement of the vessel structure. Therefore, in the following section, models trained with a U-Net architecture are proposed.

2.4 Method: U-Net

U-Net architecture

In this thesis, a U-Net structure is applied to develop a vessel segmentation model fulfilling the criteria mentioned in section 2.1. The original U-Net architecture was proposed by Ronneberger et al. in [45] and adapted to 3D input data in [6]. The architecture of the U-Net used in this work is visualized in Figure 7. It has three scales. At each scale, convolution layer and batch normalization [18] are applied, followed by ReLU (Rectified Linear Unit) activations. For better generalization, spatial dropout [60] is applied in the upstream path of the U-Net.

Implementation Details

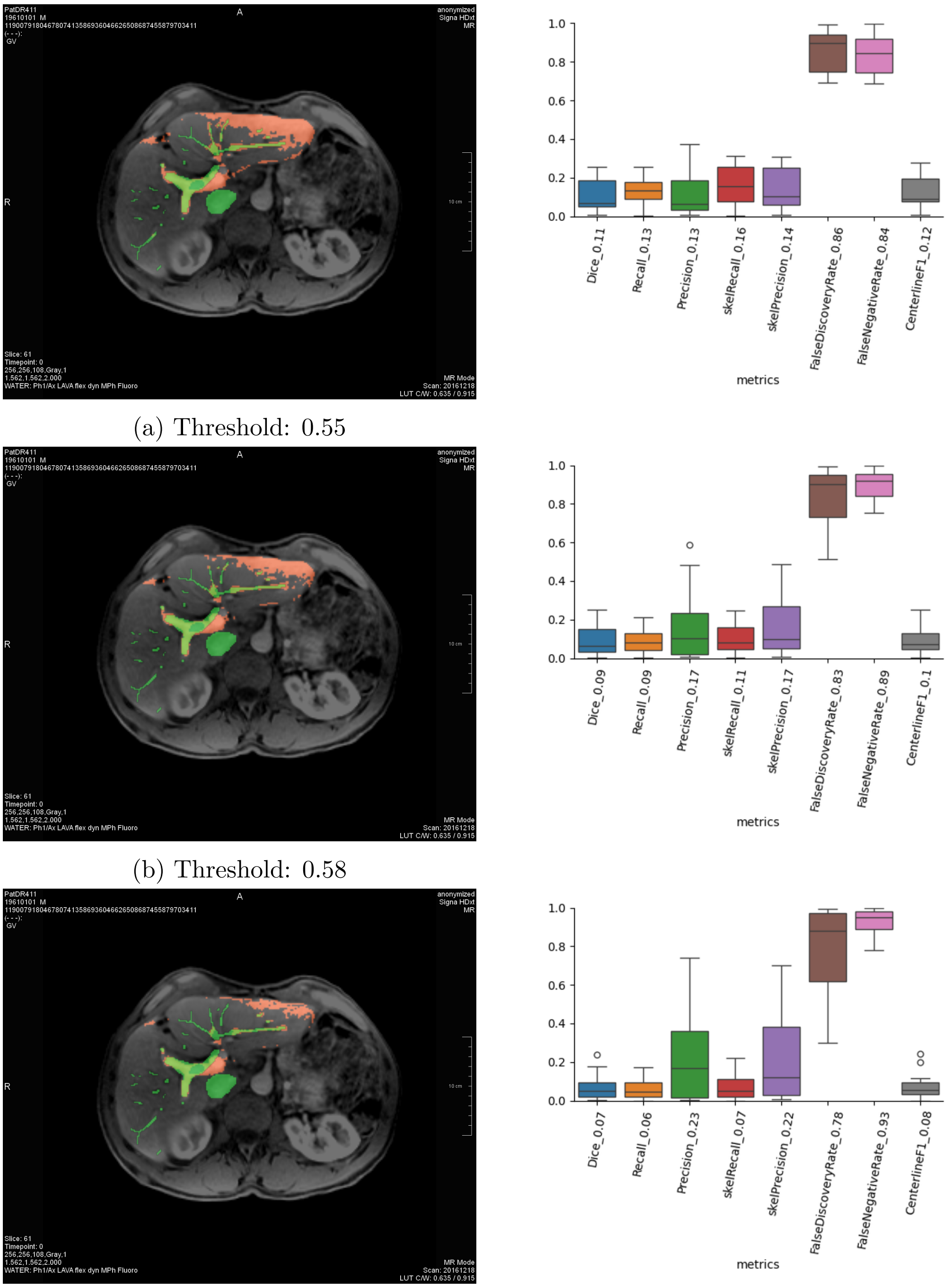
The U-Net is trained based on an implementation template by Fraunhofer MEVIS using the keras/tensorflow library. The individual loss functions are pre-defined by Fraunhofer MEVIS as well and combined in different ways with various parameter settings in this thesis. Pre- and postprocessing are implemented inside the research and development platform MeVisLab [38]. The evaluation of the models, plots of metrics and Precision-Recall curves are performed in Python.

2.4.1 Preprocessing

Data preparation

On the whole dataset, normalization and resampling is applied before further use. For normalization of the image intensities, the 2nd and 98th percentiles of the voxel value distribution are mapped to the range of 0 to 1 and afterwards, the data is clipped to the interval $[0, 1]$ to reduce artifacts.

The U-Net is trained on input data resampled to the voxel size $2\,\text{mm} \times 2\,\text{mm} \times 2\,\text{mm}$ as the original data predominantly has this resolution in the slice thickness direction



(a) Threshold: 0.55

(b) Threshold: 0.58

(c) Threshold: 0.61

Figure 6: Results of the threshold segmentation on a sample of SIRTOP data for different thresholds (left). With increasing threshold values, the number of false positives (red) decreases while the number of false negatives (green) increases. The number of true positive voxels (yellow) remains low, which shows that these segmentation results are not satisfactory. This can also be seen in the low metrics values of the box plots (right).

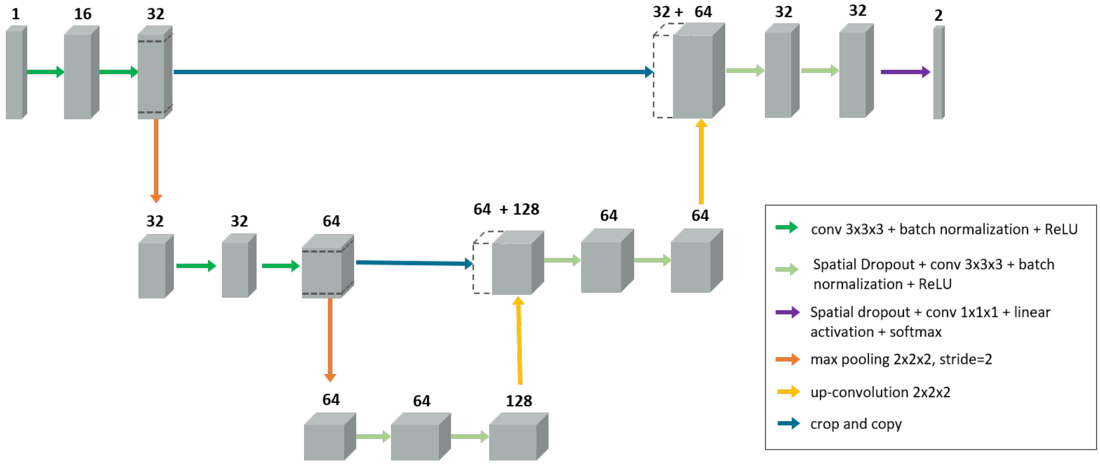

1
16
32
32 + 64
32
32
2
32
32
64
64 + 128
64
64
64
64
128
conv 3x3x3 + batch normalization + ReLU
Spatial Dropout + conv 3x3x3 + batch normalization + ReLU
Spatial dropout + conv 1x1x1 + linear activation + softmax
max pooling 2x2x2, stride=2
up-convolution 2x2x2
crop and copy

Figure 7: U-Net structure used in this work for the development of vessel segmentation models (template provided by Farina Kock and Felix Thielke).

and the chosen U-Net architecture assumes isotropic voxels for optimal performance. Resampling is done with the Lanczos3 algorithm [41]. The vessel mask is scaled to the same size with the Nearest Neighbor method to guarantee that the label values are maintained.

Data augmentation

Different augmentation techniques are applied to the patches to obtain a greater variety of training data.

Augmentation	Parameter	Prob. per patch
Rotation	rotation angle $\alpha \in [-15, 15]$	0.2
Multiplicative Brightness	multiplication factor $c_1 \in [0.75, 1.25]$	0.15
Contrast Transform	multiplication factor $c_2 \in [0.75, 1.25]$	0.15
Gamma Transform	$\gamma \in [0.7, 1.5]$	0.3
Gaussian noise	variance $\sigma_1 \in [0, 0.1]$	0.1
Gaussian blurring	variance $\sigma_2 \in [0.5, 1]$	0.2
Low Resolution Transform	sampling factor $s \in [0.5, 1]$	0.25

Table 3: Augmentation techniques and corresponding parameters applied on the training dataset. The parameters are drawn from a uniform distribution on the specified interval. The last column shows the probability of application to a patch. The augmentation methods are also used in combination.

In the following, the augmentation techniques are defined in the way they are used in this project. For the application, a MeVisLab module for data augmentation (provided by Fraunhofer MEVIS) has been used with predefined values as listed in Table 3. This module is based on the implementation and definitions in [20].

Let sample x have mean $\overline{x}$ and minimum and maximum value x_{min} and x_{max}. Then, the rotation of the image is given by

$$T_{rotation}(x, \alpha) = R_\alpha x \tag{3}$$

where R_α denotes the rotation matrix for angle α around all three axis.
For factor c_1, the transformation for multiplicative brightness is given by

$$T_{brightness}(x, c_1) = c_1 \cdot x. \tag{4}$$

For changing the contrast of the input image, the contrast transform is used:

$$T_{contrast}(x, c_2) = \mathrm{clip_x}(c_2 \cdot (x - \overline{x}) + \overline{x}) \tag{5}$$

$$\text{with } \mathrm{clip_x}(y) = \begin{cases} x_{min} & y < x_{min} \\ y & y \in [x_{min}, x_{max}] \\ x_{max} & y > x_{max} \end{cases} \tag{6}$$

where c_2 denotes the brightness factor.
The gamma transform is applied on the normalized image, which results in the following definition:

$$T_{gamma}(x, \gamma) = \left(\frac{x - x_{min}}{x_{max} - x_{min} + \epsilon} \right)^\gamma \cdot (x_{max} - x_{min} + \epsilon) + x_{min} \tag{7}$$

with $\epsilon = 1 \times 10^{-7}$. This value would not be necessary for our dataset, as it does not contain single-color images, but is implemented in this way by default.
Furthermore, the images are noised or blurred by the addition of Gaussian noise n with variance σ_1,

$$T_{noise}(x, \sigma_1) = x + n, \tag{8}$$

or the application of a Gaussian filter g_{σ_2} with kernel size $2 \cdot (\text{batchsize} \cdot \sigma_2) + 1$ and variance σ_2,

$$T_{blurring}(x, \sigma_2) = x * g_{\sigma_2}. \tag{9}$$

The Low Resolution Transformation is computed by downsampling the image with factor s with the Nearest Neighbor method and subsequently, upsampling the result to original size with bi-cubic spline interpolation.

2.4.2 Number of Labels

In section 1.3, the two largest vessel systems in the liver – hepatic and portal vein – are described. In the SIRTOP dataset, the annotations distinguish between the two vessel systems with two different labels. The models evaluated in this thesis are trained with one joint label for “vessels” as for the application no differentiation between hepatic and portal veins is necessary.

2.4.3 Loss Function

The loss function is the basis for learning the right structures in U-Nets as it measures how similar the predicted segmentation is to the reference. It is the objective function to be minimized by the optimizer and therefore decides how the network handles errors. In recent years, several loss functions for different segmentation tasks have been proposed. As described in [22, 35], the choice of the loss function can have a major impact on the results of the network and it is important to adapt it to the application. A main challenge is to account for input and output imbalance. As part of this thesis, various loss functions are analyzed and evaluated. The focus here is on the challenges mentioned in section 2.1, especially to develop a model that recognizes the long thin structures of the vessels and over-estimates the segmentation rather than under-estimating it.

Combo Loss

A main challenge in many medical segmentation tasks and especially in segmenting vessels is that the vessels are very small structures compared to the liver tissue. This results in different class distributions of the two labels as there are typically more background than foreground voxels. This is taken into account in region-based losses such as the Dice loss [62]. The *Dice loss* is defined for binary reference y and prediction $\hat{y}$ as

$$L_{Dice}(y,\hat{y}) = 1 - \frac{2\sum_{i=1}^{N} y_i \cdot \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon} \tag{10}$$

where $y_i \in \{0,1\}$ and $\hat{y}_i \in [0,1]$ denote the value of y and $\hat{y}$ in voxel i for $i = 1, ...N$. N is the number of voxels of y and $\hat{y}$ and ϵ is a small positive value to avoid division by zero [57].

This loss function measures the overlap between predicted and ground truth region. The loss function can be expressed as a fraction dependent on the number of *true positive* (TP), *false positive* (FP), and *false negative* (FN) voxels. The sum over y_i yields the number of voxels annotated as vessels in the reference which is the number of TP and FN. The sum over $\hat{y}_i$ indicates the number of voxels labeled as vessels in the prediction which is the number of TP and FP. This shows that the loss function is not affected by the number of *true negatives* (TN) which would be large in case of input imbalance.

Nevertheless, using Dice loss as loss function can lead to an output imbalance. This means that the model either segments too much or too less, because FP and FN are treated equally in Dice loss. FP increases when FN decreases and the other way round. This symmetry of FP and FN in Dice loss makes it difficult to improve the model with regard to recall or precision [59].

In distribution-based loss functions such as cross entropy loss, the error between ground truth and prediction is computed pixelwise and therefore offers a better trade-off between FP and FN. The *categorical cross entropy* (CCE) loss is defined as

$$L_{CCE}(y,\hat{y}) = -\frac{1}{N}\sum_{c\in C}\sum_{i=1}^{N} g_{i,c}\ln(p_{i,c}) \tag{11}$$

where N denotes the number of voxels of y again and C is the set containing the class labels. $g_{i,c}$ is the binary ground truth value for voxel i and class c,

$$g_{i,c} = \begin{cases} 1 & y_i = c \\ 0 & \text{else} \end{cases} \qquad \text{for all } i \in \{1, ..., N\}. \tag{12}$$

$p_{i,c}$ is the predicted "probability" for voxel i to have class label c [62].
Considering only two labels for foreground and background, the CCE loss reduces to *binary cross entropy* (BCE) [62] loss with $C = \{0, 1\}$

$$L_{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i) \right]. \tag{13}$$

While the cross entropy loss function handles the output imbalance better, it does not solve the problem of input imbalance here [59]: The sum of the pixel-wise errors depends on the number of voxels with $y_i \neq c$ for every class label $c \in C$. Therefore, if the number of labeled voxels of the background class with label $c = 0$ is significantly larger than the number of class $c = 1$, then, assuming similar probabilities, the sum $\sum_{i=1}^{N} y_i \ln(\hat{y}_i)$ is significantly larger than $\sum_{i=1}^{N} (1 - y_i) \ln(1 - \hat{y}_i)$. This results in an overrepresentation of samples where background voxels are misclassified and therefore a sensitivity to frequent class labels.

To handle both the input and output imbalance, a combination of Dice loss and BCE loss is often used in segmentation tasks and proposed as *Combo loss* in [59]. This combined loss function is defined as

$$L_{Combo}(y, \hat{y}) = \alpha L_{Dice}(y, \hat{y}) + (1 - \alpha) L_{BCE}(y, \hat{y}) \tag{14}$$

for a weighting parameter $\alpha \in (0, 1)$. In [59], an additional weighting parameter β is included in the equation (14) but has been set to 0.5 for the application in this thesis as the weighting of FP and FN is regulated with various parameters in the following section.

Focal Tversky loss
The *Tversky loss* function proposed by Salehi et al. in [46] is based on the Dice loss with additional weighting of false positive and false negative which improves output balance. It is defined as

$$L_{Tversky}(y, \hat{y}) = 1 - TI_1(y, \hat{y}) \tag{15}$$

$$\text{with } TI_c(y, \hat{y}) = \frac{\sum_{i=1}^{N} g_{i,c} p_{i,c}}{\sum_{i=1}^{N} g_{i,c} p_{i,c} + \alpha \sum_{i=1}^{N} g_{i,c} p_{i,\bar{c}} + \beta \sum_{i=1}^{N} g_{i,\bar{c}} p_{i,c} + \epsilon} \tag{16}$$

where $c = 1$ denotes the class of the vessels, TI_c is the Tversky index for class c and $\bar{c}$ denotes the opposite class of c [46]. This can solve the problem of equally treated FP and FN and allows improvement of the model with special regard to recall or precision.

Thus, the recall of the model can be improved for this application and like this, enhances the output structure to be overestimated rather than underestimated. In [1], the authors propose the use of the *Focal Tversky loss* with additional exponent γ_T:

$$L_{FocTver}(y, \hat{y}) = \sum_{c \in C} [1 - TI_c]^{\frac{1}{\gamma_T}} \tag{17}$$

for $\gamma_T \in [1, 3]$.
This improves handling of input imbalance as the parameter γ_T can equalize the influence of background class by giving more weight to hard examples of foreground class.

Focal Loss
The same principle as in the Focal Tversky loss is applied to the CCE loss in the *Focal loss* which is defined by

$$L_{Focal}(y, \hat{y}) = -\frac{1}{N} \sum_{c \in C} \sum_{i=1}^{N} (1 - p_{i,c})^{\gamma_F} g_{i,c} \ln(p_{i,c}) \tag{18}$$

proposed by Lin et al. in [33]. With the parameter γ_F, harder examples of the foreground class can be weighted more than samples from the background class. Therefore, output imbalance can be avoided.

Skeleton Recall Loss
The *Skeleton Recall loss* is proposed by Kirchhoff et al. in [27] and is defined as

$$L_{SkeletonRecall}(y, \hat{y}) = 1 - \sum_{c \in C} \frac{\sum_{i=1}^{N} \text{skel}(g_{i,c}) \cdot p_{i,c}}{\sum_{i=1}^{N} \text{skel}(g_{i,c})}. \tag{19}$$

skel$(\cdot)$ describes the operation of determining the skeleton of the structure. This is realized by a MeVisLab module for skeletonization which is based on the publication by Selle et al. [49]. The authors describe the method as a successively symmetrical erosion of the surface voxels in such a way that the topology of the vessel structure is preserved. The Skeleton Recall metric measures how much of the skeleton of the annotated vessel structure is predicted as vessel in the model output. This loss function gives more weight to the vessel structure itself and therefore, simplifies learning of long thin structures as vessels.

Tested Loss Functions
Based on the above-mentioned advantages and disadvantages of the individual loss functions, segmentation models with the following loss functions are trained as part of this Master's thesis.

1. The **Combo loss** is mainly used in existing models [21, 59] and therefore, has been evaluated in this thesis to be comparable to these models.

2. The **Tversky loss** has been chosen as comparison to the Combo loss as it can be considered as generalization of the Dice loss with improved output imbalance.

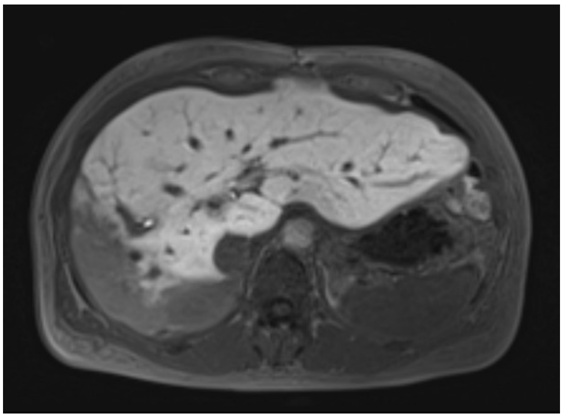


Figure 8: In this slice of an MRI volume of the hepatobiliary phase of a PSC patient, varying intensity values within the liver are visible, especially in the lower left.

3. The combination of **Focal Tversky loss + Focal loss** is considered as an alternative to the Combo loss, where both Dice loss and BCE are generalized and can be evaluated with different weighting parameters. This loss function is also denoted as Hybrid Focal loss in [62].

4. The **Skeleton Recall loss** has been evaluated in combination with **Focal Tversky loss** and **Focal loss** here. In existing models, it is often combined with the Combo loss [27]. However, as the combination of Focal Tversky loss and Focal loss show better results for this application, the Skeleton Recall loss is combined with these in this thesis.

All loss functions are evaluated with different parameters in section 2.5.1.

2.4.4 Global Intensity Non-linear Augmentation

A striking feature in MRI images of the late phase of PSC patients are varying intensity values within the liver tissue [3] as visualized in Figure 8. To account for this in the vessel segmentation model developed in this work, the application of the *global intensity non-linear* (GIN) augmentation method presented in [44] is evaluated. GIN augmentation transforms the input images into images with different intensity values to imitate various acquisition methods and to pay more attention to the content, i.e., the shapes of the anatomical structures, during training.
This is done via a shallow 3D convolutional network with the following structure:

1. The image of the training dataset on which GIN augmentation is performed is denoted as input x.

2. A shallow multi-layer convolutional network is applied with a number of hidden channels and hidden layers, each containing

a. convolution with random kernels θ selected from Gaussian distribution $\mathcal{N}(0, I)$

b. application of Leaky ReLU $f(x) = \max(0.1 \cdot x, x)$ as non-linearity.

In the following, the output of the convolutional network is denoted as $g_\theta^{Net}(x)$.

3. The original image x and the output $g_\theta^{Net}(x)$ are combined linearly:

$$\tilde{g}_\theta(x) = \alpha g_\theta^{Net}(x) + (1 - \alpha)x.$$

4. Finally, the output is re-normalized to have the same Frobenius norm as the original input:

$$g_\theta(x) = \frac{\tilde{g}_\theta(x)}{\|\tilde{g}_\theta(x)\|_F} \cdot \|x\|_F.$$

In this work, the MeVisLab module for GIN augmentation (provided by Fraunhofer MEVIS) has been used with 2 hidden channels and 4 hidden layers in the convolutional network. The kernel size is chosen as 1 in every layer, as there is a risk that the narrow vascular structures will disappear when convolving with a larger kernel. While a convolution with a kernel size of 1 is typically viewed as performing channel-wise scaling, the application of LeakyReLU introduces nonlinear transformations that can produce outputs resembling different modalities. Therefore, the GIN augmentation method remains effective even with this kernel size.

Another advantage of GIN augmentation is that a larger amount of data could be used for training. Due to the independence of acquisition processes that the network learns, CT images are suitable training data as well. Therefore, in this work, both MRI and CT images of the SIRTOP dataset are used for training the models in which GIN augmentation is used. GIN augmentation is applied to 50% of the training and validation data. The test dataset is the same as described in section 2.4.6 still containing only MRI data, because the application is specifically designed for this modality and to allow for a meaningful comparison with other models.

2.4.5 N3: Nonparametric Nonuniform Intensity Normalization

The nonuniformity of signal intensities in MRI data appears due to inhomogeneity in the main magnetic field and can result in $10 - 20\%$ variations in image intensities. This has little impact on visual assessment but can significantly influence the results of automatic DL segmentation models [52].

Different methods exist to reduce intensity nonuniformity in MRI data, but the *nonparametric method for automatic correction of intensity nonuniformity* (N3) proposed by Sled et al. in [52] neither requires longer acquisition time nor further hardware, e.g., for measuring the radio frequency field, which makes it universally applicable.
For N3, the nonuniformity is modeled as a smooth multiplicative field f in

$$v(x) = u(x) \cdot f(x) + n(x) \tag{20}$$

where v denotes the measured signal, u the true signal and n is white Gaussian noise assumed to be independent of u. In the noise-free case, the equation simplifies with $\hat{u}(x) := \log(u(x))$ to an additive model

$$\hat{v}(x) = \hat{u}(x) + \hat{f}(x). \tag{21}$$

Under the assumption that $\hat{u}$ and $\hat{f}$ are independent or uncorrelated random variables, the distribution V of $\hat{v} = \hat{u} + \hat{f}$ is given by

$$V(\hat{v}) = (F * U)(\hat{v}) \tag{22}$$

with U and F the probability densities of $\hat{u}$ and $\hat{f}$. This characterizes the nonuniformity distribution F as blurring of the true intensity distribution U. Based on these theoretical assumptions, the authors in [52] develop the N3 method which alternatingly estimates the nonuniformity field $\hat{f}$ and the distribution U. The space of possible fields $\hat{f}$ is restricted to be smooth and slowly varying and therefore, based on the evaluation of different MRI data, the distribution F is assumed to be approximately Gaussian. F is approached by deconvolution of Gaussian distributions from estimated U in each iteration. In this way, the entire space of possible distributions for U corresponding to Gaussian distributions of F can be searched as every Gaussian distribution can be described as convolution of two Gaussians. The method consists of the two following alternating iteration steps:

1. The first step is the estimation of the field $\hat{f}$ for given distribution U which is realized by

   $$\hat{f}_s(\hat{v}) = S\{\hat{v} - E[\hat{u}|\hat{v}]\} \tag{23}$$

   where the expected value $E[\hat{u}|\hat{v}]$ of $\hat{u}$ given $\hat{v}$ can be computed based on the distributions U and F and the measured signal $\hat{v}$ which is explained in more detail in [52]. The subtraction $\hat{v} - E[\hat{u}|\hat{v}]$ is computed at each location x and then, a smoothing operator S is applied to remove high-frequency components, as $\hat{f}$ is assumed to be smooth and slowly varying.

2. The second step is the estimation of the distribution U. For this, the convolution in equation 22 is considered in the Fourier space, becoming a multiplication

   $$\mathcal{F}(V) = \mathcal{F}(F)\mathcal{F}(U) \tag{24}$$

   where $\mathcal{F}(\cdot)$ denotes the Fourier transform. Then, U can be estimated pointwise in Fourier space with the following deconvolution filter

   $$\mathcal{F}(U) = \frac{\mathcal{F}(F)^*}{|\mathcal{F}(F)|^2 + Z^2}\mathcal{F}(V) \tag{25}$$

   where * denotes the complex conjugate and Z a constant term for limitation of the magnitude.

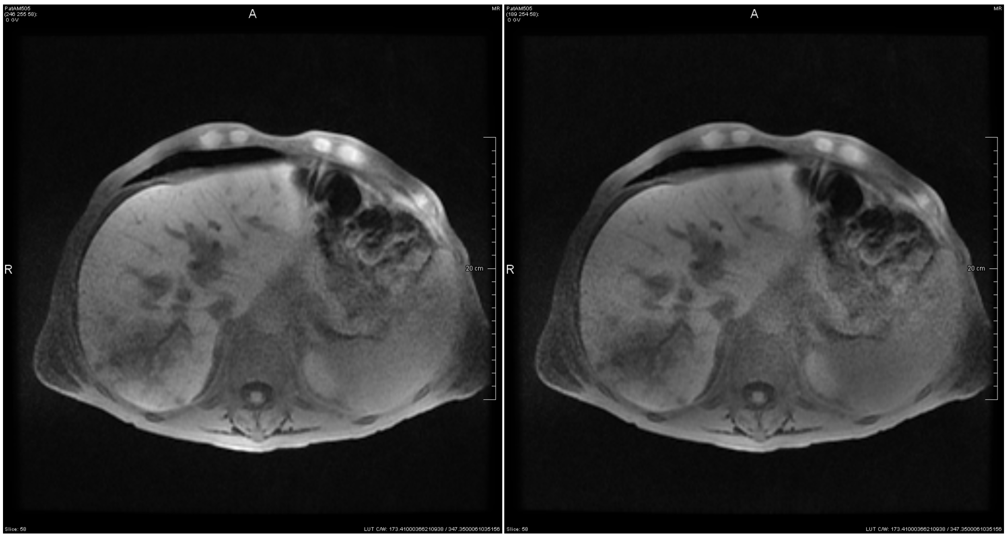


Figure 9: Original MRI image of a patient from the SIRTOP dataset (left) and results after application of N3 (right), resulting in less inhomogeneity of the intensity values.

The distribution V that is required for the computations in step 2 is estimated by a triangular Parzen window based on histogram values with equal-size bins. To reduce computational costs, the MRI data v is resampled to lower resolution.

In this work, a MeVisLab module for N3 (provided by Fraunhofer MEVIS) has been used for the implementation. The subsampling factor has been chosen as 2. For histogram sharpening, the number of bins has been selected as 200 with a width of 0.15 for the deconvolution kernel and a regularization of 0.01.

The N3 method offers the possibility to be applied on a masked area of the MRI volume which could lead to even more compensation of the nonuniformity. Nevertheless, in this work, the method is without mask as it was not yet certain how well the liver segmentation algorithm would run on the PSC data in particular and therefore, without liver mask, the vessel segmentation model is independent of the accuracy of any liver segmentation model. Figure 9 shows an MR image before and after the application of N3.

2.4.6 Training

To train the U-Net, the set of MRI volumes of SIRTOP data described in section 1.5.2 is split into 46 samples for training, 8 samples for validation and 15 for test. The volumes are divided into approximately 400.000 patches of size 128 × 128 × 128 for training and approximately 100.000 patches of size 52 × 52 × 52 with zero-padding 20 × 20 × 20 for validation. The batch size is chosen as 4 and the number of iterations as 25.000. In the following, the optimizer and the learning rate used for training are described in more detail.

Optimizer

For the training of the U-Net, the NovoGrad algorithm is used as optimizer. NovoGrad is an optimization algorithm based on stochastic gradient descent, which was proposed by Ginsburg et al. in [12] and especially developed for the training of deep networks. The idea of this optimization method is to combine the advantages of Adam [26] and *Stochastic Gradient Descent* (SGD) with momentum [58] to create a method that performs well for different DL tasks and is more robust to the choice of the initial learning rate [12]. The authors show in [12] that NovoGrad performs equal or better than popular algorithms such as SGD with momentum, Adam and AdamW. The main difference to SGD and Adam is that the algorithm uses layer-wise gradient normalization to be more robust to noisy gradients. This idea has been proposed in several papers, e.g., [63, 51] and in case of NovoGrad is combined with the Adam method.

For each layer, new weights w_{t+1} are computed based on modified first moments $\mathbf{m}_t^l$ and second moments v_t^l as known from the Adam algorithm. In iteration t for layer l, the following computation steps are performed:

$$v_t^l = \beta_2 \cdot v_{t-1}^l + (1-\beta_2) \cdot \|\mathbf{g}_t^l\|^2, \tag{26}$$

$$\mathbf{m}_t^l = \beta_1 \cdot \mathbf{m}_{t-1}^l + \left(\frac{\mathbf{g}_t^l}{\sqrt{v_t^l + \epsilon}} + d \cdot \mathbf{w}_t^l \right), \tag{27}$$

$$\mathbf{w}_{t+1}^l = \mathbf{w}_t^l - \lambda_t \cdot \mathbf{m}_t. \tag{28}$$

The variable g_t^l contains the gradient $\mathbf{g}_t^l = \nabla_l L(\mathbf{w}_t)$ of the loss function, λ_t is the learning rate and β_1, β_2 and d are weighting parameters. In this work, the implementation from the tensorflow library has been used with standard parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-7}$ and $d = 0.0$. The learning rate λ_t choice is described in the following paragraph.

Learning Rate

As learning rate, polynomial decay has been applied which is defined by

$$\lambda_i = \lambda_0 \cdot (1 - \frac{i}{N})^p \tag{29}$$

for $i = 1, ..., N$ where N denotes the number of iterations. λ_0 is the initial learning rate and p the power [39]. In this thesis, the power is chosen as $p = 0.9$ since this value has been used in [34, 4] and showed good performance regarding training duration and model accuracy. The number of iterations is $N = 25000$. For λ_0, two different values are tested, $\lambda_0 \in \{0.01, 0.1\}$, but no major impact on the result is observed. In Figure 10, the loss plot is compared for $\lambda_0 = 0.01$ and $\lambda_0 = 0.1$ for the Combo loss function. It can be seen that the resulting training and validation loss is nearly the same for both values, although for $\lambda_0 = 0.1$ the loss curve is more noisy at the beginning. The evaluation on the test data indicates nearly the same results as shown in the comparison plots of various metrics between reference and prediction in Figure 11. The mean metrics values are the same apart from a deviation of 0.01 in the recall value and there are

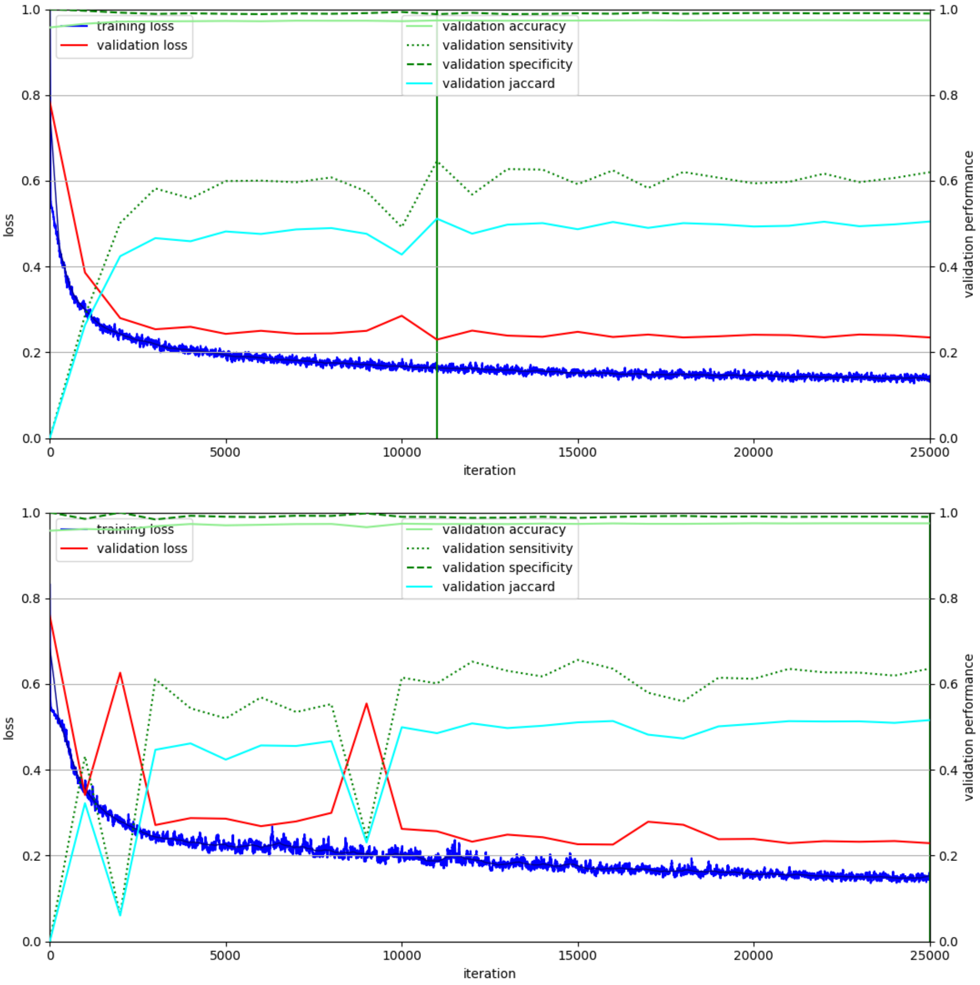


Figure 10: Loss evolution for learning rate $\lambda_0 = 0.01$ (above) and $\lambda_0 = 0.1$ (below) using the Combo loss. The evolution for $\lambda_0 = 0.1$ shows strong variations in training and validation loss, while for $\lambda_0 = 0.1$ the evolutions are much smoother. Nevertheless, the resulting values after 25000 iterations are nearly the same.

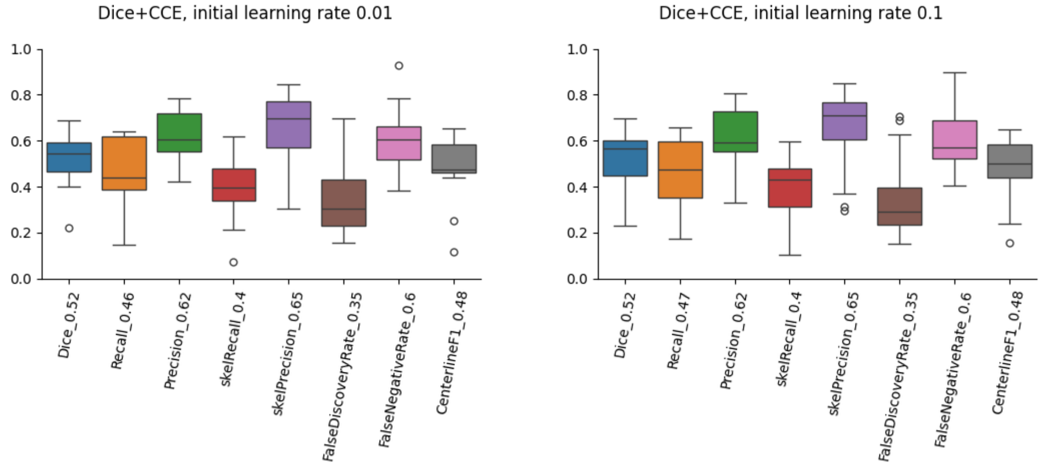


Figure 11: The evaluation results of the models trained with the Combo loss for learning rate $\lambda_0 = 0.01$ (left) and $\lambda_0 = 0.1$ (right) show only minor differences in the variance and median values of the box plots. The mean metrics values (listed with underscore in the labels) are – apart from a minimal deviation in the recall value – the same (metrics defined in section 2.5.1).

only minor differences in the variance and median values over the test dataset. In [12], the robustness of NovoGrad to the initial learning rate is confirmed. Therefore, in the following, λ_0 is chosen as 0.01 for all trainings.

Threshold method for uncalibrated output

In [47], it is shown that deep learning models can have inferior calibrated probabilities as output under certain conditions such as application of Dice loss or batch normalization. In [30], the authors propose a temperature calibration to avoid this problem. However, in this thesis, the simple method of varying the threshold after the softmax function is evaluated. This threshold is typically at 0.5: voxels with output values above this threshold are labeled as vessels and below as background. As in this application high recall is required, a lower threshold than 0.5 is chosen to receive more voxels labeled as vessels.

In section 2.6, different thresholds are applied and studied based on Precision-Recall curves and a suitable threshold is chosen for the clinical evaluation of the models.

2.5 Evaluation of Vessel Segmentation Models

2.5.1 Evaluation on Test Data

In this thesis, a U-Net with 9 loss function parameter settings is trained on the SIRTOP dataset presented in section 1.5.2. The dataset is preprocessed in three different ways: the MRI dataset with preprocessing methods as described in section 2.4.1, the same dataset with additionally N3 applied, and the MRI dataset combined with CT data and GIN augmentation applied. This results in a total of 27 models as listed in Table 4. For the evaluation on the test dataset, for each model eight different metrics are computed. The selection of suitable metrics is presented in the following paragraph. Based on the results on test data, the best models are selected for a clinical evaluation performed by the medical student of this project on the dataset for liver function computation. This evaluation is presented in section 2.5.2.

Metrics for Evaluation

The choice of the metrics for evaluation of segmentation results depends on the task. For the evaluation on the test dataset, the following metrics are considered.
A widely used metric for semantic segmentation tasks is the DSC [9]. It measures the overlap between reference y and prediction $\hat{y}$ and is defined as

$$\mathrm{DSC}(y, \hat{y}) = \frac{2 \cdot |y \cap \hat{y}|}{|y| + |\hat{y}|} \tag{30}$$

where $|y|$ denotes the number of voxels in y with value 1 (analog for $\hat{y}$) and $|y \cap \hat{y}|$ the number of voxels with value 1 in both volumes x and y [36].
The DSC is based on the two metrics for measuring the precision and the recall [36] that are defined as

$$\mathrm{T}_{prec}(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|\hat{y}|} \tag{31}$$

$$\mathrm{T}_{recall}(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y|}. \tag{32}$$

T_{prec} evaluates what fraction of the voxels that are predicted as vessels are in fact annotated as vessels in the ground truth. T_{recall} indicates what fraction of the voxels labeled as vessels in the reference are predicted as vessels.
However, in the presence of very small structures as given in case of vessel segmentation, DSC, recall and precision can be inadequate [36, 50] as thin structures have less impact on the metrics in comparison to larger regions.

In [50], the authors propose a variant of DSC for measuring the overlap between two tubular-shaped structures, the *centerlineDice* (clDice). It is based on the *Topology Precision* metric T_{clPrec} (skelPrecision) and the *Topology Sensitivity* metric $\mathrm{T}_{skelRecall}$ (skelRecall) which are computed based on the skeleton of the reference ($s(y)$) and predicted segmentation ($s(\hat{y})$):

$$\mathrm{T}_{clPrec}(y,\hat{y}) = \frac{|y \cap s(\hat{y})|}{|s(\hat{y})|} \tag{33}$$

$$\mathrm{T}_{skelRecall}(y,\hat{y}) = \frac{|s(y) \cap \hat{y}|}{|s(y)|}. \tag{34}$$

With T_{clPrec} and $T_{skelRecall}$, clDice is defined as

$$\mathrm{clDice}(y,\hat{y}) = \frac{2 \cdot \mathrm{T}_{clPrec}(y,\hat{y}) \cdot \mathrm{T}_{skelRecall}(y,\hat{y})}{\mathrm{T}_{clPrec}(y,\hat{y}) + \mathrm{T}_{skelRecall}(y,\hat{y})} \tag{35}$$

and $\mathrm{clDice}(y,\hat{y}) \coloneqq 0$ if $\mathrm{T}_{clPrec}(y,\hat{y}) + \mathrm{T}_{skelRecall}(y,\hat{y}) = 0$ [50].
In addition, *False Discovery Rate* (FDR) and *False Negative Rate* (FNR) are considered in the context of skeleton metrics. They are defined as

$$\mathrm{FDR}(y,\hat{y}) = 1 - \mathrm{T}_{clPrec}(y,\hat{y}) \tag{36}$$

$$\mathrm{FNR}(y,\hat{y}) = 1 - \mathrm{T}_{skelRecall}(y,\hat{y}). \tag{37}$$

As mentioned above, for the segmentation of vessels, it can be more adequate to consider skeleton metrics, as they give more weight to small structures. However, in this application, the evaluation of the metrics without skeleton – recall, precision and DSC – can still be interesting as not only the main vascular structure and branching pattern but also the boundary areas of the vessels are relevant. Therefore, in the following all metrics defined above are considered.

For the selection of the best model, the focus is on the recall and skelRecall metrics as they express how many of the (skeleton) voxels that are annotated as vessels are predicted to be vessels and as explained in section 2.1, overestimation is preferred in the application of this project. In the skelRecall, small thin structures are given more weights. In the recall metric, the accuracy near the boundary of the segmentation is also considered.

Impact of preprocessing methods
The model with the highest recall value – Model 7a – is trained on the dataset preprocessed with GIN augmentation. In general, it is observed that GIN augmentation increases the recall values in all models where Focal Tversky loss and Focal loss are included in the loss function. As these loss functions have already larger recall values in comparison to models 1 and 2, GIN augmentation seems to strengthen the effect of high recall values.

However, increasing recall values come with decreasing precision values. The plot in Figure 12 corresponding to the model with the largest recall shows a very low precision of 0.05 and the segmentation on the MRI image in Figure 13 indicates that the predicted segmentation contains a lot of tissue outside of the liver. This phenomenon appears in several models in this work especially in those with GIN augmentation combined with the Skeleton Recall loss contained in loss function (Model 6a to 9a) and the one with higher weighting of FP (larger value for β_T) in Tversky loss (Model 4a).

Model	Lossfunction	Parameter	GIN	N3	Liver	Dice	Recall	Precision	skelRec	skelPrec	FDR	FNR	clDice
1a	dice+cce	a=0.5, b=0.5	✓	×	×	0.49	0.45	0.54	0.38	0.57	0.43	0.62	0.45
1a	dice+cce	a=0.5, b=0.5	✓	×	✓	0.53	0.51	0.59	0.38	*0.65*	*0.35*	0.62	0.47
1b	dice+cce	a=0.5, b=0.5	×	✓	×	0.51	0.46	0.6	0.41	0.63	0.37	0.59	0.49
1b	dice+cce	a=0.5, b=0.5	×	✓	✓	0.55	0.51	0.6	0.41	*0.65*	*0.35*	0.59	0.49
1c	dice+cce	a=0.5, b=0.5	×	×	×	0.52	0.46	*0.62*	0.4	*0.65*	*0.35*	0.6	0.48
1c	dice+cce	a=0.5, b=0.5	×	×	✓	*0.56*	0.52	**0.63**	0.41	**0.66**	**0.34**	0.59	0.49
2a	tversky	α_T = **0.3**, β_T = **0.7**, γ_T = **1**	✓	×	×	0.49	0.55	0.46	0.47	0.53	0.47	0.53	0.48
2a	tversky	α_T = **0.3**, β_T = **0.7**, γ_T = **1**	✓	×	✓	0.52	0.59	0.49	0.47	0.58	0.42	0.53	0.5
2b	tversky	α_T = **0.3**, β_T = **0.7**, γ_T = **1**	×	✓	×	0.53	0.57	0.51	0.52	0.58	0.42	0.48	0.53
2b	tversky	α_T = **0.3**, β_T = **0.7**, γ_T = **1**	×	✓	✓	*0.56*	0.62	0.52	0.52	0.59	0.41	0.48	0.54
2c	tversky	α_T = **0.3**, β_T = **0.7**, γ_T = **1**	×	×	×	0.54	0.55	0.54	0.47	0.6	0.4	0.53	0.52
2c	tversky	α_T = **0.3**, β_T = **0.7**, γ_T = **1**	×	×	✓	**0.57**	0.6	0.55	0.47	0.61	0.39	0.53	0.52
3a	tversky+focal	a=0.5, b=0.5, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	✓	×	×	0.51	0.58	0.46	0.49	0.5	0.5	0.51	0.48
3a	tversky+focal	a=0.5, b=0.5, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	✓	×	✓	0.55	0.62	0.5	0.49	0.58	0.42	0.51	0.52
3b	tversky+focal	a=0.5, b=0.5, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	✓	×	0.53	0.56	0.53	0.51	0.59	0.41	0.49	0.54
3b	tversky+focal	a=0.5, b=0.5, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	✓	✓	*0.56*	0.6	0.54	0.51	0.61	0.39	0.49	0.54
3c	tversky+focal	a=0.5, b=0.5, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	×	×	0.52	0.54	0.52	0.49	0.57	0.43	0.51	0.52
3c	tversky+focal	a=0.5, b=0.5, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	×	✓	0.55	0.59	0.53	0.49	0.59	0.41	0.51	0.52
4a	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	✓	×	×	0.3	0.7	0.2	0.64	0.18	0.82	0.36	0.27
4a	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	✓	×	✓	0.49	0.73	0.37	0.64	0.47	0.53	0.36	0.53
4b	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	✓	×	0.51	0.67	0.41	0.63	0.51	0.49	0.37	0.55
4b	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	✓	✓	0.53	0.71	0.43	0.63	0.53	0.47	0.37	**0.57**
4c	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	×	×	0.5	0.66	0.41	0.61	0.5	0.5	0.39	0.54
4c	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	×	✓	0.52	0.71	0.42	0.61	0.53	0.47	0.39	*0.56*
5a	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **3**, γ_F=2	✓	×	×	0.42	0.71	0.3	0.65	0.35	0.65	0.35	0.44
5a	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **3**, γ_F=2	✓	×	✓	0.48	0.74	0.37	0.65	0.45	0.55	0.35	0.52
5b	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **3**, γ_F=2	×	✓	×	0.5	0.67	0.41	0.62	0.52	0.48	0.38	*0.56*
5b	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **3**, γ_F=2	×	✓	✓	0.52	0.71	0.42	0.62	0.54	0.46	0.38	**0.57**
5c	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **3**, γ_F=2	×	×	×	0.5	0.65	0.42	0.6	0.5	0.5	0.4	0.53
5c	tversky+focal	a=0.5, b=0.5, α_T = **0.1**, β_T = **0.9**, γ_T = **3**, γ_F=2	×	×	✓	0.52	0.69	0.43	0.6	0.53	0.47	0.4	0.55
6a	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	✓	×	×	0.19	0.72	0.12	0.73	0.1	0.9	0.27	0.17
6a	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	✓	×	✓	0.43	0.77	0.31	0.73	0.37	0.63	0.27	0.47
6b	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	✓	×	0.48	0.65	0.39	0.66	0.43	0.57	0.34	0.52
6b	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	✓	✓	0.52	0.71	0.42	0.67	0.47	0.53	0.33	0.54
6c	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	×	×	0.48	0.64	0.4	0.63	0.44	0.56	0.37	0.5
6c	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	×	✓	0.51	0.69	0.42	0.63	0.47	0.53	0.37	0.52
7a	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	✓	×	×	0.09	*0.78*	0.05	*0.77*	0.06	0.94	*0.23*	0.1
7a	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	✓	×	✓	0.37	**0.82**	0.24	**0.78**	0.31	0.69	**0.22**	0.43
7b	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	✓	×	0.44	0.73	0.32	0.73	0.4	0.6	0.27	0.51
7b	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	✓	✓	0.46	*0.78*	0.34	0.73	0.43	0.57	0.27	0.54
7c	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	×	×	0.43	0.71	0.32	0.68	0.39	0.61	0.32	0.49
7c	tversky+focal+skeleton	a=**0.3**, b=**0.3**, c=**0.4**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	×	✓	0.46	0.76	0.34	0.69	0.44	0.56	0.31	0.52
8a	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	✓	×	×	0.3	0.67	0.2	0.66	0.18	0.82	0.34	0.26
8a	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	✓	×	✓	0.49	0.72	0.39	0.66	0.43	0.57	0.34	0.51
8b	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	✓	×	0.52	0.62	0.46	0.62	0.49	0.51	0.38	0.53
8b	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	✓	✓	0.55	0.67	0.47	0.62	0.51	0.49	0.38	0.55
8c	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	×	×	0.51	0.6	0.45	0.59	0.48	0.52	0.41	0.52
8c	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.3**, β_T = **0.7**, γ_T = **1**, γ_F=2	×	×	✓	0.54	0.65	0.47	0.59	0.5	0.5	0.41	0.53
9a	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	✓	×	×	0.16	0.74	0.1	0.72	0.11	0.89	0.28	0.17
9a	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	✓	×	✓	0.43	*0.78*	0.3	0.73	0.39	0.61	0.27	0.49
9b	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	✓	×	0.48	0.68	0.38	0.67	0.46	0.54	0.33	0.54
9b	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	✓	✓	0.5	0.73	0.39	0.67	0.48	0.52	0.33	0.55
9c	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	×	×	0.47	0.68	0.37	0.65	0.46	0.54	0.35	0.52
9c	tversky+focal+skeleton	a=**0.4**, b=**0.4**, c=**0.2**, α_T = **0.1**, β_T = **0.9**, γ_T = **2**, γ_F=2	×	×	✓	0.5	0.72	0.38	0.65	0.48	0.52	0.35	0.54

Table 4: Overview of all evaluated models. For each selected loss function and parameter combination, three models are trained on a differently preprocessed dataset and evaluated with and without application of a liver mask. The best values in each column are marked in red showing that Model 7a has the highest recall values. Thus, this model is best suited for the application considered in this thesis.

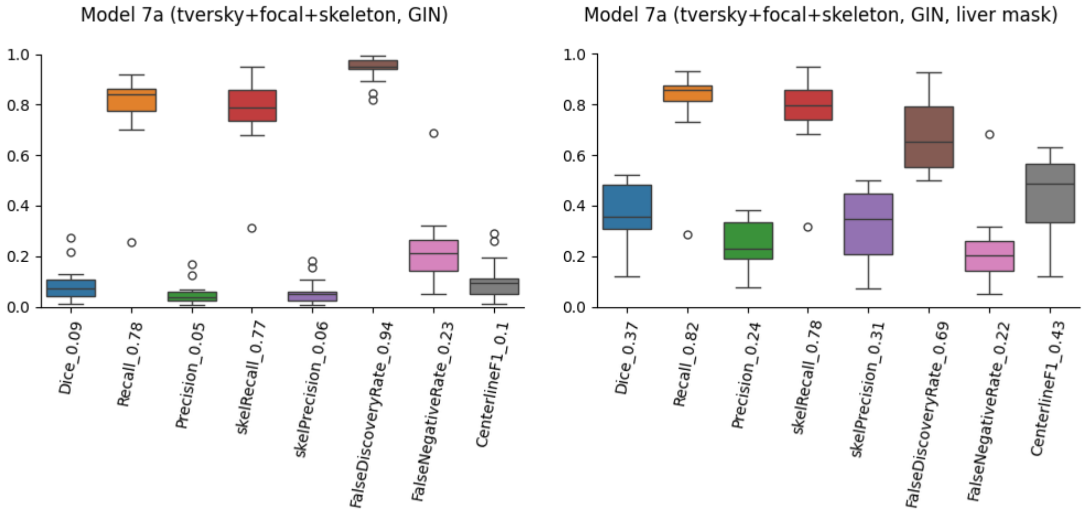


Figure 12: Model 7a shows very low precision and skelPrecision values as visualized in the boxplot on the left. Therefore, a liver mask is applied resulting in larger precision values as can be seen on the right.

It seems like there has not been enough background in training. But – although the patches are chosen as “foreground patches” which means they are only included for training if they contain structures annotated as vessels – the patch size is large enough to include background voxels in the patches. Additionally, the fact that this phenomenon only appears in context of GIN augmentation in combination with certain loss functions rather suggests that it is related to the augmentation method or the usage of CT data.

Since in this project the aim is to segment pure liver parenchyma and therefore, liver segmentation is required in the automatic workflow, the liver segmentation model [15] can be used for masking the predicted vessel segmentation and in this way improving the precision for the models mentioned above. For reasons of comparable results, the liver mask is applied to all model results and evaluated additionally to the predicted output. The mask is chosen as the largest connected component of the union over the liver segmentation and the predicted vessel segmentation. Thus, it contains the region of the liver and additionally all regions that are predicted as vessels and connected to the mask of the liver segmentation. This choice aims to include the vessels in the marginal area of the liver segmentation to compensate for any inaccuracies in the liver segmentation. The results for the masked vessel segmentation are shown in Table 4 with an appropriate sign in column “liver”. In model 7a, the precision value increases to 0.24 in case of application of the liver mask. The skelRecall value is 0.82, then, and the value of the recall is 0.78 which is even higher than without liver mask.

The N3 method also improves the recall values but not as much as with GIN augmentation. The application of N3 increases the recall in the average by only 0.03 but actually improves the values for all loss functions except the Combo loss.

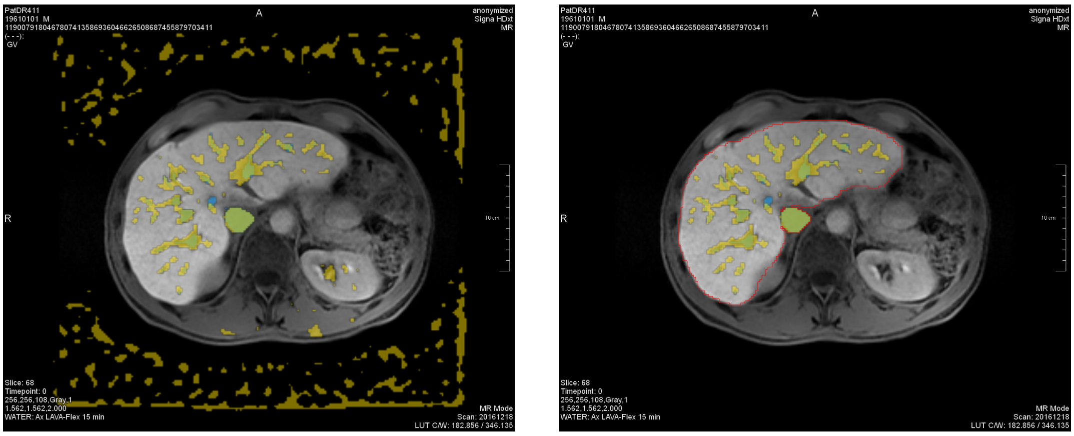


Figure 13: Model 7a segments many erroneous background structures outside of the liver (left). With application of a liver mask (red contour), the segmentation is reduced to the region of the liver (right). FP is yellow, TP is green and FN is blue.

Overall, the two preprocessing methods – GIN augmentation and N3 – improve the vessel segmentation models with regard to the application in this project. Therefore, the models based on these methods are preferred in the selection of the best models.

Impact of the Loss Function

As already mentioned in section 2.4.3, the choice of the loss function can have a high impact on the model output. This is confirmed by considering the evaluation results.

The models trained with the Combo loss, which is often applied for vessel segmentation, show the highest precision and skelPrecision values in comparison to the other models evaluated within this thesis. They reach mean precision values of 0.59 (without liver mask applied) and 0.61 (with liver mask) in comparison to 0.5 (without liver mask) and 0.52 (with liver mask) or below in the other models. The skelPrecision value is also larger with mean values of 0.62 (with liver mask) and 0.65 (without liver mask) in the models with Combo loss as loss function and mean values of 0.57 and 0.59 or below in the other models. This difference is not surprising, as in this work the loss function is improved based on the Combo loss in order to increase the recall.

For the Focal Tversky loss function, parameters $\alpha_T = 0.3$ and $\beta_T = 0.7$ are chosen first. This increases the recall values by 0.10 in the mean and precision decreases in the mean by 0.08 in comparison to the model with Combo loss. The application of Focal loss changes these values only marginally. However, increasing the weight for FP to $\beta_T = 0.9$ (and decreasing $\alpha_T = 0.1$) and choosing the exponent γ_T as 2 or 3 increases the recall values considerably.

The addition of Skeleton Recall loss to the combination of Focal Tversky and Focal loss increases the recall values even more. This is to be expected, as the definition of this loss function is based on the measurement of the skelRecall metric. The recall and skelRecall values of the models increase with a higher weighting of the Skeleton Recall loss function. At the same time, precision decreases. This can be observed in the comparison of models 7a-c to models 9a-c where Focal Tversky loss, Focal loss and Skeleton Recall loss are weighted with 0.3, 0.3 and 0.4 vs. 0.4, 0.4 and 0.2.

These results show that we could adapt the vessel segmentation model step by step to the given task. We improved the loss function starting with Combo loss and changing it with several weighting parameters and additional losses resulting in a combination of loss functions which shows large recall values. In particular, the weighting of FP and FN as well as the addition of Skeleton Recall loss improve the model the most. Thus, these models have great potential for the application in the automatic workflow and are preferred in the model selection presented in the next paragraph.

Model selection

In Table 5, the model selection for the clinical evaluation is presented. In the following, the selection criteria are explained.

	Model	Selection criterion
1	7a	largest recall and skelRecall value
2	7b	largest recall and skelRecall without GIN augmentation
3	4b	largest clDice
4	4c	largest clDice without N3
5	9a	direct comparison of all three preprocessing methods, relatively high recall and precision at the same time
6	9b	
7	9c	
8	5b	largest clDice (same as model 4b)

Table 5: This table gives an overview over the models selected for the clinical evaluation.

The model with the highest skelRecall (0.77) and the largest recall value (0.78) is model 7a with the combination of Focal Tversky loss, Focal loss and Skeleton Recall loss as loss function and GIN augmentation applied in preprocessing. To reduce the effect of segmenting structures outside of the liver segmentation described above, the model with the highest recall among all models without GIN augmentation is identified. This is model 7b with a skelRecall value of 0.73 and a recall value of 0.73 without liver mask and 0.78 with liver mask. For this model, N3 is applied on data before training and as loss function the combination of Focal Tversky loss, Focal loss and Skeleton Recall loss is used as in model 7a.

In addition to the recall, the clDice is taken into account in the model selection to evaluate whether larger precision values are required on PSC and control group data. The models with the largest clDice are models 4b and 5b. In both models, N3 is applied as preprocessing. For better comparison of the different preprocessing methods, model 4c is selected as well as model with the largest clDice among all models without N3.

Furthermore, the models 9a-c are selected to have a direct comparison of training on all three datasets and since these models show relatively high recall and precision values at the same time. In the following, the selected models are numbered from 1 to 8 as in the first column of Table 5.

Comparison with existing models
In this paragraph, the models proposed in this thesis are compared to existing models for vessel segmentation on MRI, specifically to the models discussed in section 2.2.

The highest DSC scores of the proposed models are achieved by model 2c (with liver mask) with a value of 0.57 and model 1c, 2b and 3b with a value of 0.56. These are approximately in the same range as the DSC scores of the model proposed by Zbinden et al. in [64] with values of 0.634 for pv and 0.532 for hv. For model 1c, even the precision values are very close to the ones in [64]. The Precision-Recall curves visualized in the paper by Zbinden et al. show almost linearly decreasing precision values for increasing recall. Nevertheless, there are no exact values for recall given, making the comparison more difficult. Furthermore, based on the arguments above, the models with higher recall such as model 7a or 7b are selected for further use in this thesis.

Compared to the model proposed by Oh et al. in [43], the models in this thesis deviate more strongly in DSC values. However, there are again no recall values given in the work by Oh et al. as the aim of this application is different and does not focus on over-estimation. For the kind of application presented in this thesis, no published model is found. The model proposed by Zhao et al. in [65] shows the best DSC, precision and recall values with 0.8724, 0.8776 and 0.8648. All metrics values are presented in Table 6 for direct comparison.

Model	DSC score	Recall	Precision
nnU-Net [64]	0.634/0.532 (pv/hv)	-	-
Residual U-Net [43]	0.61/0.70/0.58 (pv/hv/bd)	-	-
TTGA U-Net [65]	0.8724	0.8776	0.8648
Model 1c (ours)	0.56	0.52	0.63
Model 2b (ours)	0.56	0.62	0.52
Model 2c (ours)	0.57	0.6	0.55
Model 3b (ours)	0.56	0.6	0.54
Model 7a (ours)	0.37	0.82	0.24
Model 7b (ours)	0.46	0.78	0.34

Table 6: Comparison of DSC scores of existing vessel segmentation models. The TTGA U-Net shows the best DSC scores among the models for MR images.

To sum up, the models proposed in this thesis show slightly worse DSC scores compared to existing vessel segmentation models. However, the comparison of the more relevant precision and recall values is difficult because of missing values and different application tasks. The model by Zhao et al. demonstrates a considerably better balance between recall and precision and therefore shows ways that can potentially lead to better results. It would be interesting to train and test our models and the model by Zhao et al. on the same dataset. The results of the direct comparison could help to further improve our models, but this is beyond the scope of this thesis. Instead, a clinical evaluation in context of the given application is done. This enables detailed feedback on the models providing ideas for further improvement.

2.5.2 Evaluation on PSC and Control Group Data

The eight models selected based on the criteria mentioned above are evaluated by the medical student of this project based on the rating scale in Table 7. This scale focuses on the question whether the vessel segmentation can be used for the estimation of liver function. Therefore, it is less significant if too much tissue is segmented. Instead, the segmentation is rated poorly if parts of the vessel structure are not recognized.

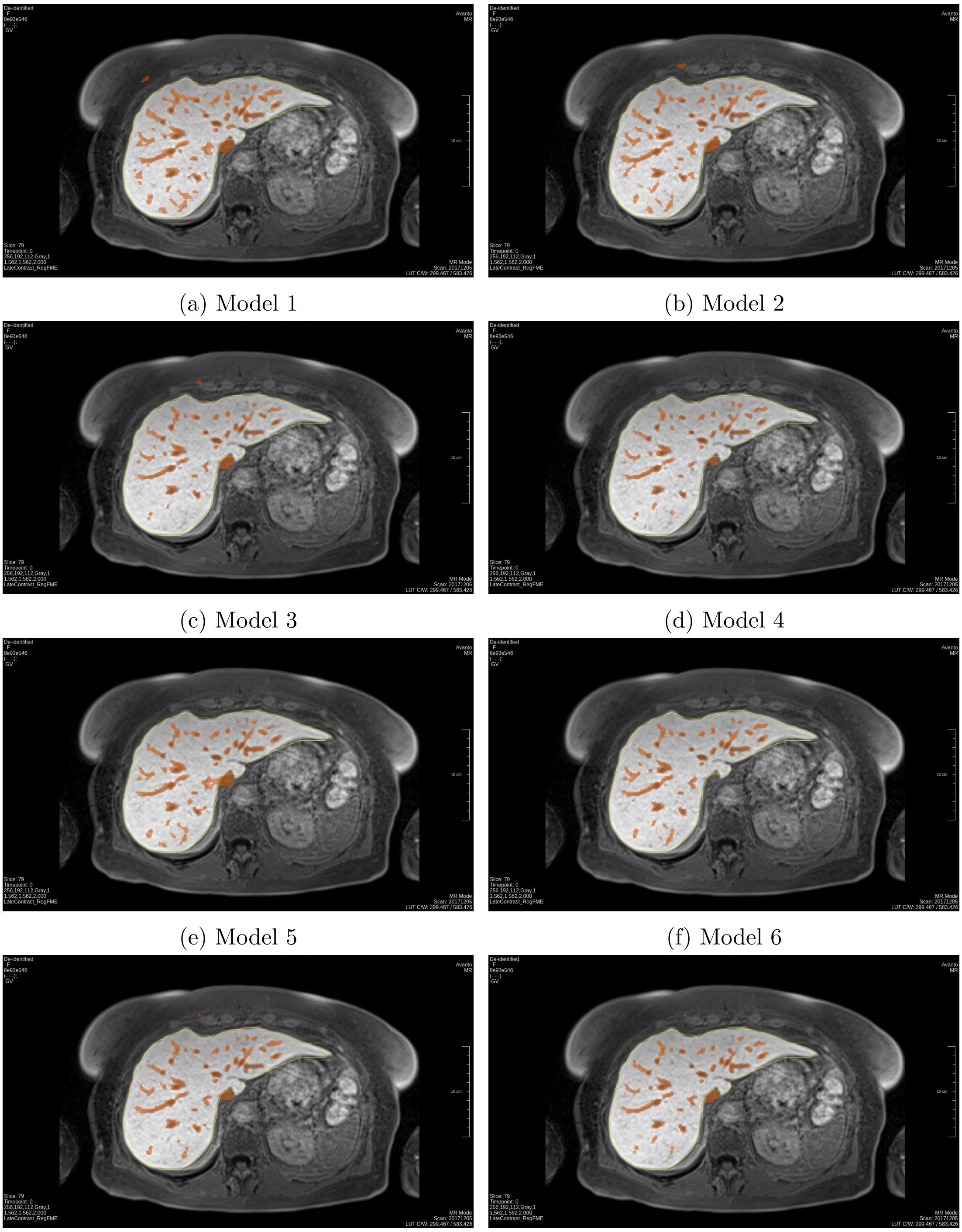
0	poor segmentation results not usable or only partially usable – unreliable RE values
1	fair segmentation, many/large errors results generally usable – relevant influence on RE values possible
2	good segmentation, some (small) errors results usable and generally accurate – irrelevant influence on RE values assumed
3	very good segmentation, minimal or no errors results very accurate – reliable RE values

Table 7: This rating scale is used for the evaluation on PSC and control group data.

The evaluation is performed on a dataset of 10 PSC patients and 10 patients of the control group. The results are presented in Table 8 and show that models 1, 2 and 5 perform best as those have the largest evaluation scores. Nevertheless, there is scope for further improvement especially for the segmentation on PSC data containing some “0”-ratings. Overall, the main vessels are detected very good but abnormalities and varying intensities especially in MR images of PSC patients are difficult to segment leading to “0” or “1”-ratings. Therefore, results on PSC data are worse than those of the control group. The cases rated with score 2 mainly contain the whole vessel structure with some inaccuracies on the boundary of the vessels or small vessels with bright intensities that can hardly be distinguished from liver tissue are not detected. In the following, we take a closer look at possible sources of error.

First of all, for some cases, all eight models provide already good results as they are mainly rated with score 2. This is visualized in Figure 14. It can be seen that, although the differences in segmentations are very small, models 1, 2 and 5 have the largest volume of segmented vessels, as they also recognize very thin vessels that are barely distinguishable from the tissue. This fits with the previous observation that these models show the largest recall values in the evaluation on test data. However, all models segment the main vessel structures without any gaps or inaccuracies in the marginal region.

In contrast to the control group, on the dataset of PSC patients, an underestimation occurs in many cases resulting in bad rating results due to parts of the vessels not being recognized as such. This is visualized in Figure 15a, for example, where dilated bile ducts are not segmented, or even stronger in the case presented in Figure 15b. Here, liver parenchyma is very inhomogeneous and leads to large parts of the vessels not being segmented. These models are trained on datasets preprocessed with GIN augmentation



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

(e) Model 5

(f) Model 6

(g) Model 7

(h) Model 8

Figure 14: An example case of the control group, where the eight selected models show very good results segmenting all main structures of the vessels. There are only small differences in the segmentations as models 1, 2 and 5 recognized slightly more structures and are therefore best suited for this application.

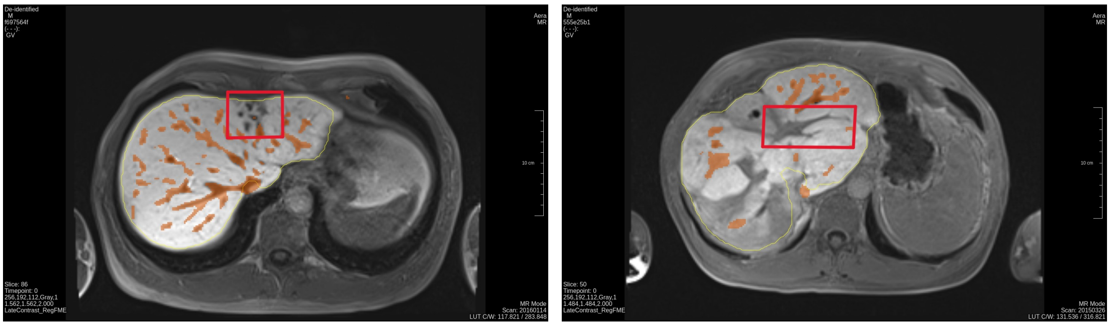
Model	**0**	**1**	**2**	**3**	**Total**
1 (7a)	4	4	2	-	8
2 (7b)	7	2	1	-	4
3 (4c)	8	-	2	-	4
4 (4b)	8	-	2	-	4
5 (9a)	6	2	2	-	6
6 (9b)	8	2	-	-	2
7 (9c)	7	3	-	-	3
8 (5b)	10	-	-	-	0

(a) Evaluation on PSC Data

Model	**0**	**1**	**2**	**3**	**Total**
1	-	4	6	-	16
2	1	3	6	-	15
3	6	3	1	-	5
4	6	4	-	-	4
5	3	4	3	-	10
6	4	4	2	-	8
7	6	3	1	-	5
8	6	3	1	-	5

(b) Evaluation on Control Group Data

Table 8: The eight best vessel segmentation models are clinically evaluated. These tables show the number of ratings with value "0" to "3" for each model separated into evaluation on PSC and control group data. The number in "Total" is computed by summing up all scores for the corresponding model.



(a) Model 1

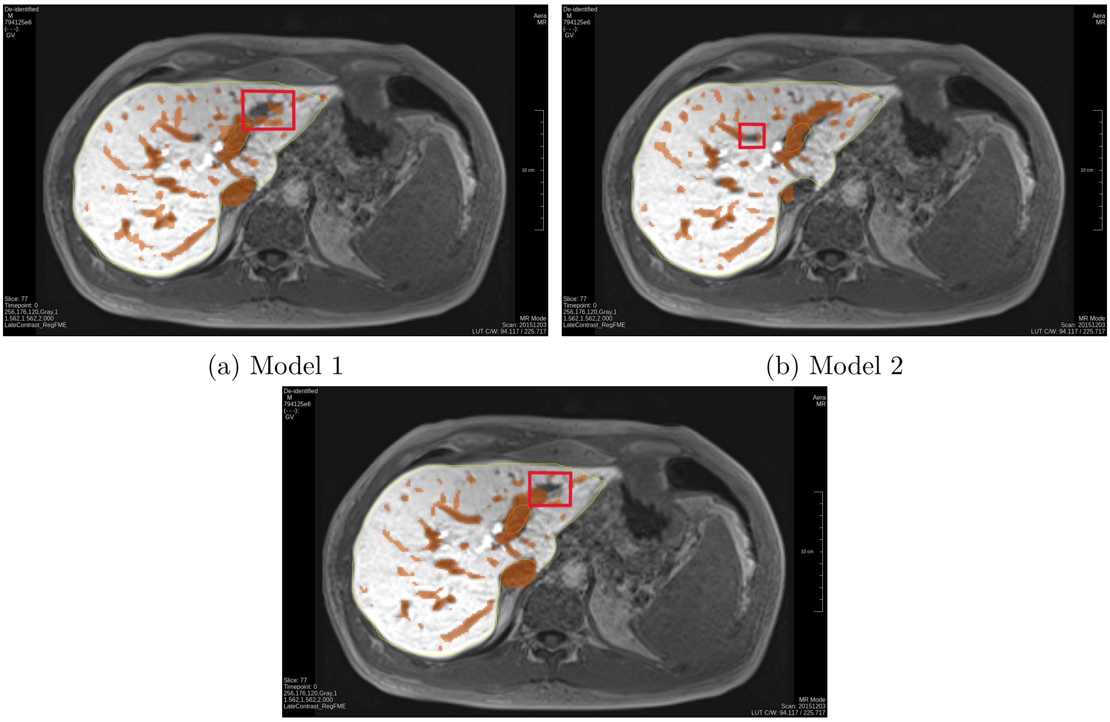
(b) Model 1

Figure 15: In this figure, two cases of underestimation are shown. On the left, abnormal bile ducts are not segmented and on the right, large vessels are not recognized due to the strong varying intensities within the slice.

and N3 to reduce the dependency on inhomogeneity in data, however, the model is not adapted sufficiently to this type of inhomogeneity with strongly varying intensities within a slice.

In Figure 16, another case of underestimation is presented. Here, parts of the largest vessels are not detected. This error appears mainly in the results of the segmentation models with GIN augmentation while the models with N3 segment the whole vessel. In contrast, the segmentation result of model 2 (with N3) contains gaps in long thin structures which are segmented better by model 1 and 5 (with GIN augmentation). Therefore, in the next section, the combination of the different models as improved model for vessel segmentation is evaluated.

Another error appearing in Figure 16 is that marginal areas are not accurately segmented everywhere. This is probably due to the fact that the annotations on SIRTOP



(a) Model 1

(b) Model 2

(c) Model 5

Figure 16: Different forms of underestimation. While model 1 and model 5 show problems in segmenting large vessels completely, the segmentation of model 2 shows gaps in narrower vessel structures.

data partly contain similar inaccuracies as can be seen in Figure 17. To enhance the models with regard to this error, a threshold method is applied in the next section segmenting more voxels as vessels to improve the error of underestimation especially in marginal areas.

Beside the error of underestimation, also some cases of overestimation appear as visualized in Figure 18. In 18a, parts of a lesion are segmented and in 18b, a ligamentary structure is partly recognized as “vessel”. These structures are no vessels but should be subtracted from liver segmentation in this application anyway. Therefore, in the following, the focus is kept on improving underestimation, as these errors have more impact on the resulting liver function values.

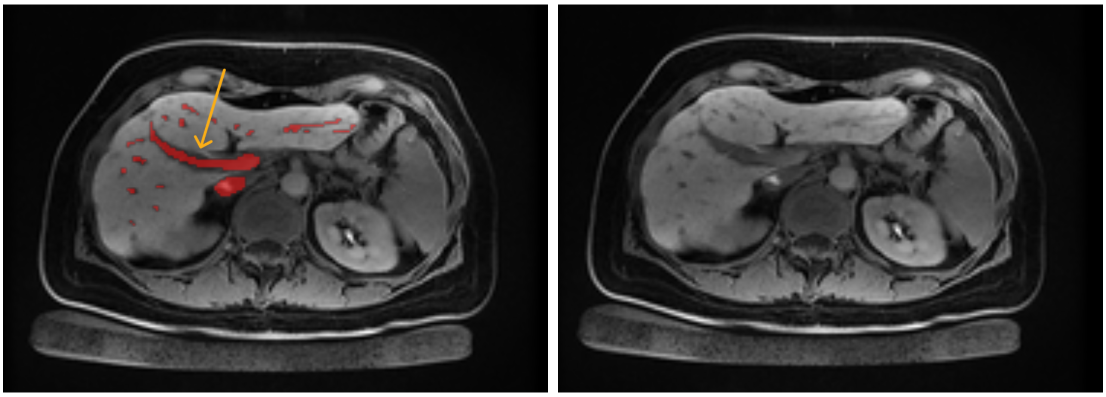
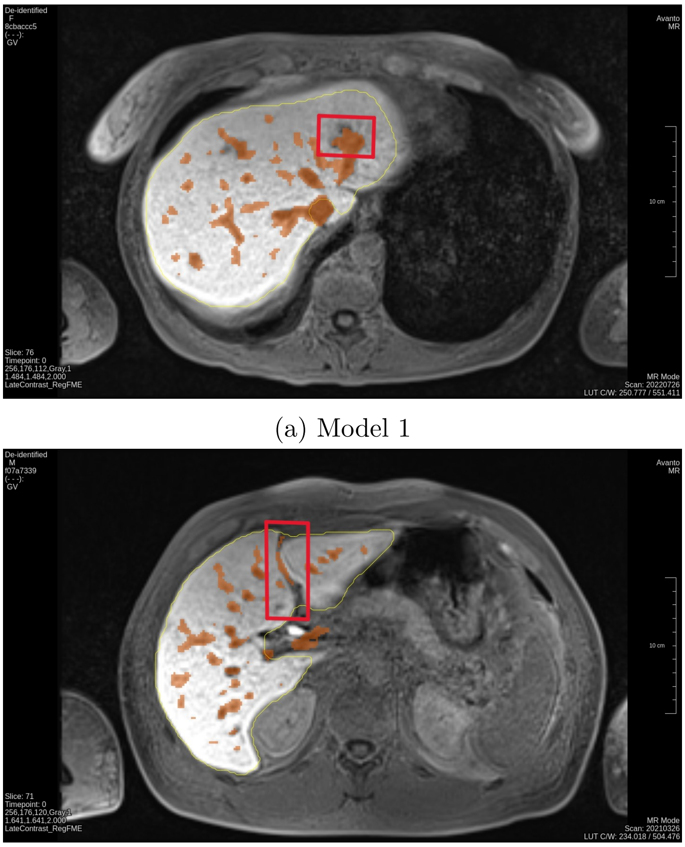


Figure 17: In a few cases, the vessel annotations of SIRTOP data (left) show marginal inaccuracies in the boundary area of the vessels. This can be seen in comparison with the corresponding slice without annotations (right).



(a) Model 1

(b) Model 1

Figure 18: These two slices show forms of overestimation. In the case above, a lesion is partly segmented and below, the segmentation contains parts of a ligamentary structure.

2.6 Improvement Methods

In this work, two strategies for improvement are proposed and applied on the models 1, 2 and 5 evaluated to be the best in section 2.5.2. It is investigated whether the two approaches can improve the previous results, both on the test data with direct comparison to the true annotations and in the clinical evaluation.

Threshold Method for Calibration

As explained in the last paragraph of section 2.4.6, a threshold method is applied on the models within this work for better calibration of the output.

To evaluate the impact of changing the threshold in ML models, the *receiver operating characteristic* (ROC) curve is widely used [10, 11]. The ROC curves for all models of Table 4 are visualized in Figure 19. The curves indicate the fraction of vessels truly annotated as vessels – the true positive rate (TPR) – and the fraction of background voxels erroneously labeled as vessels – the false positive rate (FPR) for changing threshold [11]. Different ROC curves can be compared by measuring the area under the curve (AUC) [11]. For our models, the AUC values are very large indicating a good performance as the optimal AUC value is 1. However, in the ROC curves, the class imbalance between vessels and background is not taken into account as voxels that are erroneously labeled as vessels are considered as fraction of all background voxels leading to overoptimistic results.

Therefore, the *Precision-Recall* (PR) curve is computed based on the test dataset to determine the best threshold. This is a method to visualize the trade-off between recall and precision of a model [66]. Figure 20 shows the PR curves for all models of Table 4 again, comparing the application of GIN augmentation, N3 and “None” (indicating that neither N3 nor GIN augmentation is applied). The value in brackets indicates the *average precision* (AP) which is a metric taking into account recall and precision value. It is defined as the area under the PR curve:

$$\mathrm{AP}(p) = \int_0^1 p(r)dr \tag{38}$$

where $p(r)$ denotes the PR curve as function assigning a precision to each recall value [66]. In the implementation, the discrete version is computed.

For optimal trade-off between precision and recall, the AP value would be 1. For N3 and “None”, the AP is slightly larger than 0.5 for all models except Model 2. Considering the plot of Model 7a-c evaluated to be the best in the clinical evaluation, the slope of the curve changes for the approximate recall value of 0.8. For higher recall values, the precision is decreasing faster, especially for the N3 method.

For GIN augmentation, the AP is below the values of N3 and “None” as these plots are created based on the vessel segmentation without liver mask. The application of the liver mask increases the precision for higher recall values as can be seen in Figure 21. Here, all three preprocessing methods show very similar results, with GIN augmentation showing the best precision values for the largest recall values. This confirms the previous evaluation results.

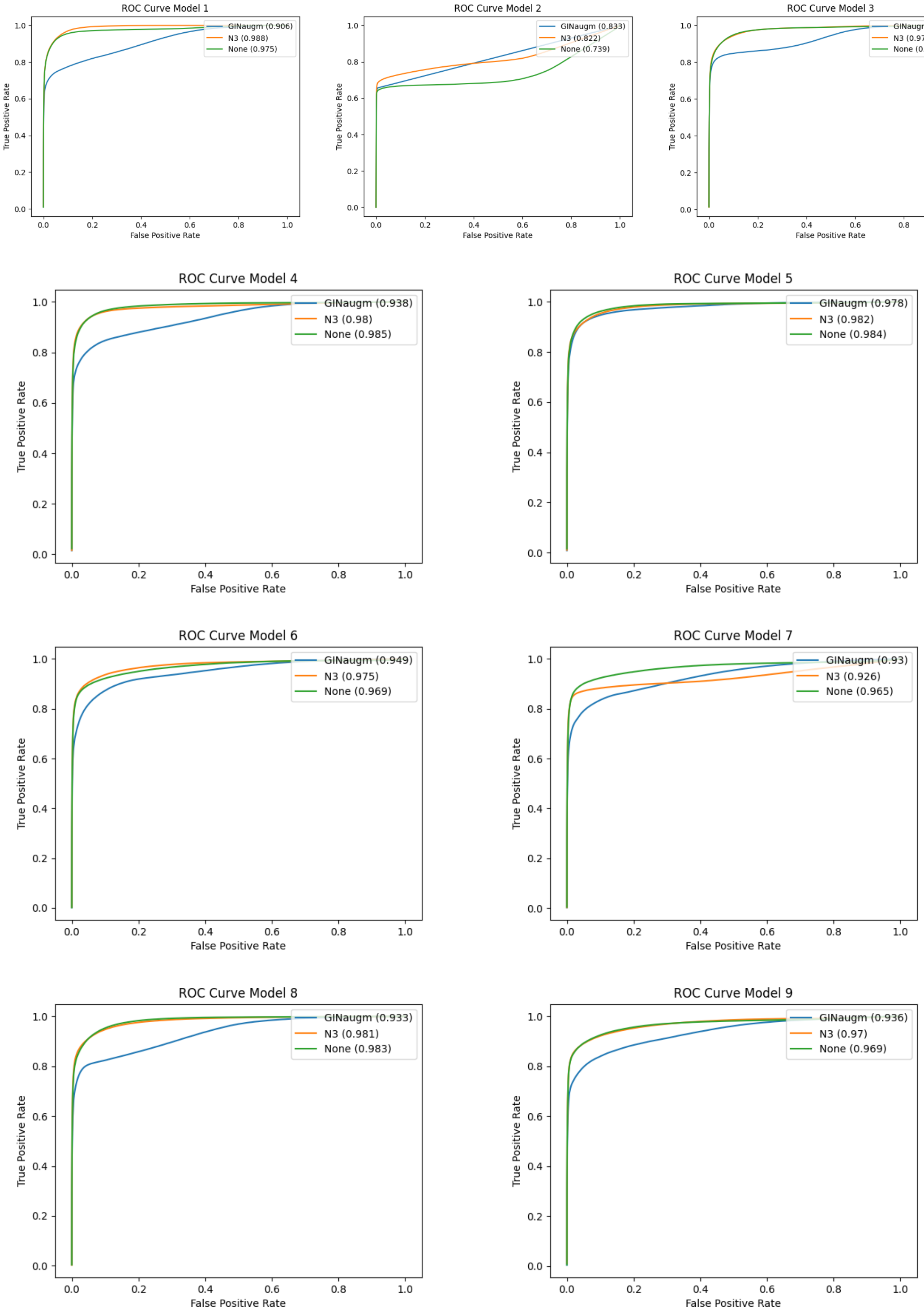


Figure 19: The mean ROC curves over all test cases for all models in Table 4. The AUC values are indicated in brackets showing good results for all models. However, these results can be overoptimistic in case of class imbalance as given in this application.

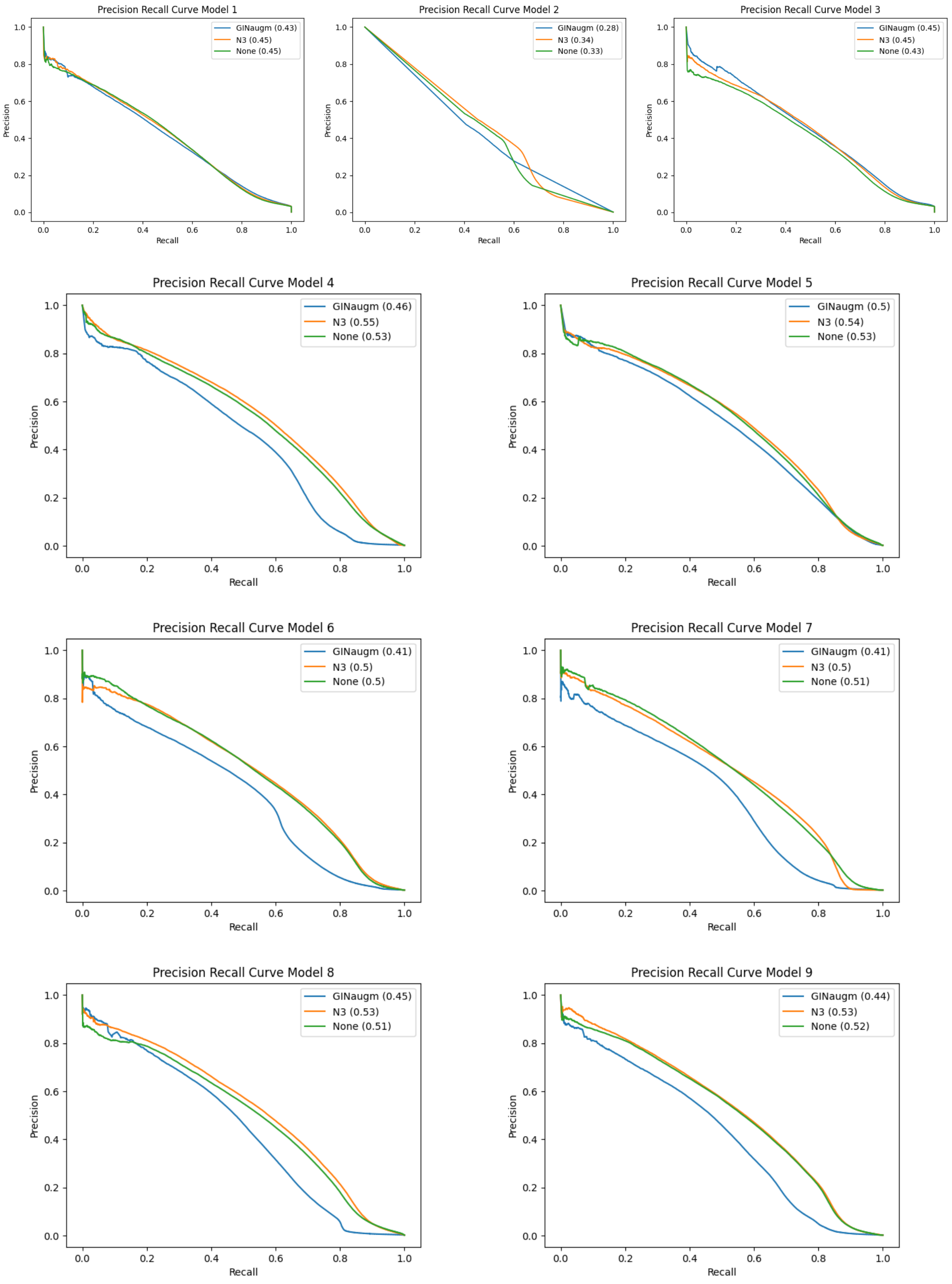
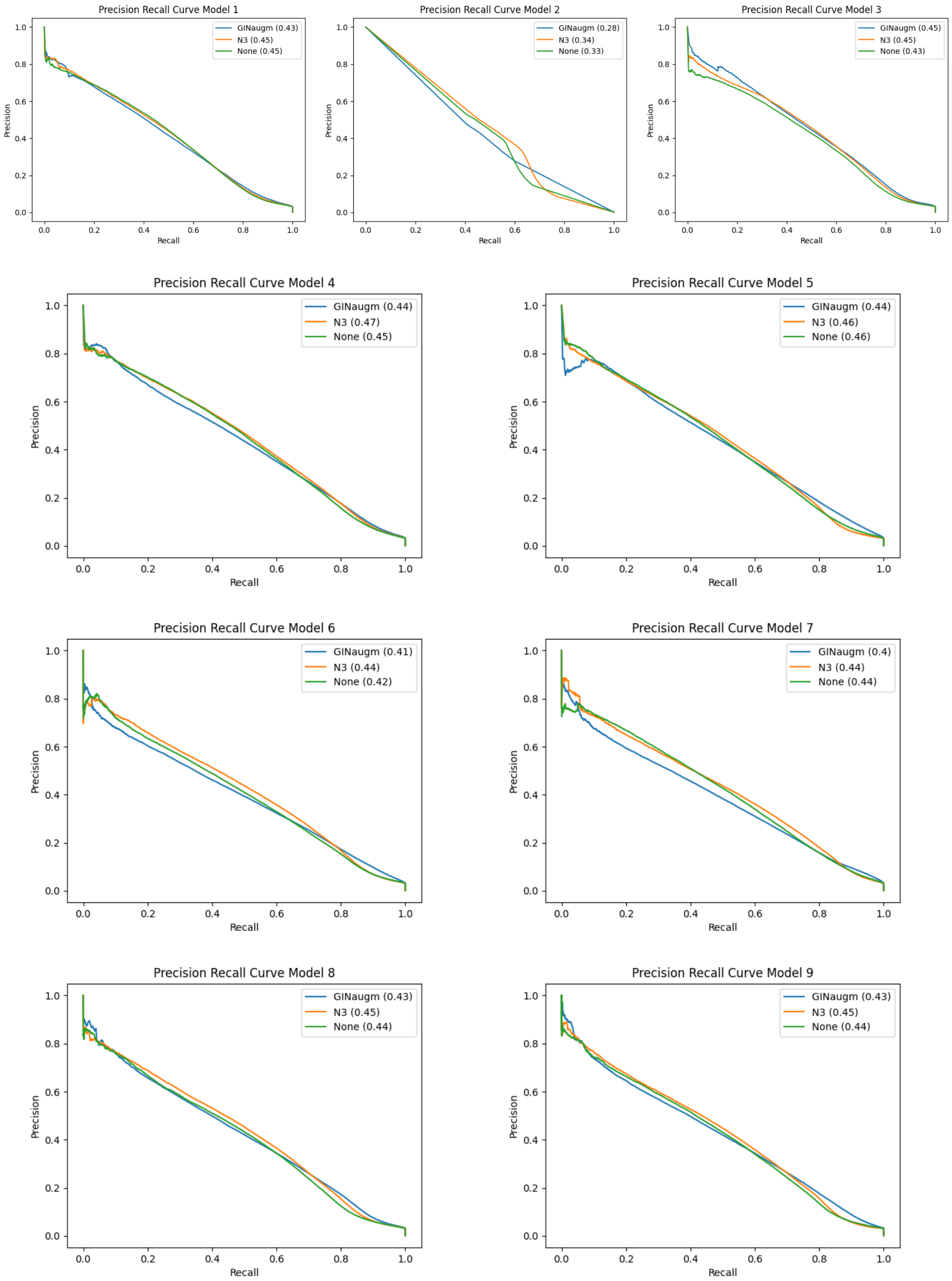


Figure 20: The mean PR curves over all test cases with the AP values indicated in brackets. The models with “N3” or “None” preprocessing outperform those with GIN augmentation due to the phenomenon of low precision values described in section 2.5.1.


Precision Recall Curve Model 1

Precision Recall Curve Model 2

Precision Recall Curve Model 3

Precision Recall Curve Model 4

Precision Recall Curve Model 5

Precision Recall Curve Model 6

Precision Recall Curve Model 7

Precision Recall Curve Model 8

Precision Recall Curve Model 9

Figure 21: The mean PR curves for all models with masked prediction and reference with the AP values indicated in brackets. All plots show approximately linear curves, making it difficult to choose a suitable threshold for model calibration.

The PR curves with liver mask are approximately linear making it difficult to choose a suitable threshold. To nevertheless assess the effects of a change in the threshold, 0.1 is chosen as threshold value for the clinical evaluation. This results in an increase of the recall values of the models selected to be the best in the clinical evaluation as visualized in the box plots in Figure 22.

As an outlook of this work, it would be interesting to find a way to increase the AP value and thus achieve a higher precision with high recall at the same time. Ideas for this are discussed in section 4.

The threshold method also improves the results of the clinical evaluation as presented in Table 9. The improved models segment more voxels especially in the marginal regions of the vessels. Furthermore, they improve the segmentation of larger vessels. Thus, the error of underestimation described in section 2.5.2 is reduced. However, at the same time more liver tissue is segmented by the improved models, so that a further reduction of the threshold is probably not useful.

Model	**0**	**1**	**2**	**3**	**Total**
1	4	4	2	-	8
1_impr	4	1	5	-	11
2	7	2	1	-	4
2_impr	5	3	2	-	7
5	6	2	2	-	6
5_impr	5	2	1	2	10

(a) Evaluation on PSC Data

Model	**0**	**1**	**2**	**3**	**Total**
1	-	4	6	-	16
1_impr	-	3	6	1	18
2	1	3	6	-	15
2_impr	1	2	7	-	16
5	3	4	3	-	10
5_impr	3	3	4	-	11

(b) Evaluation on Control Group Data

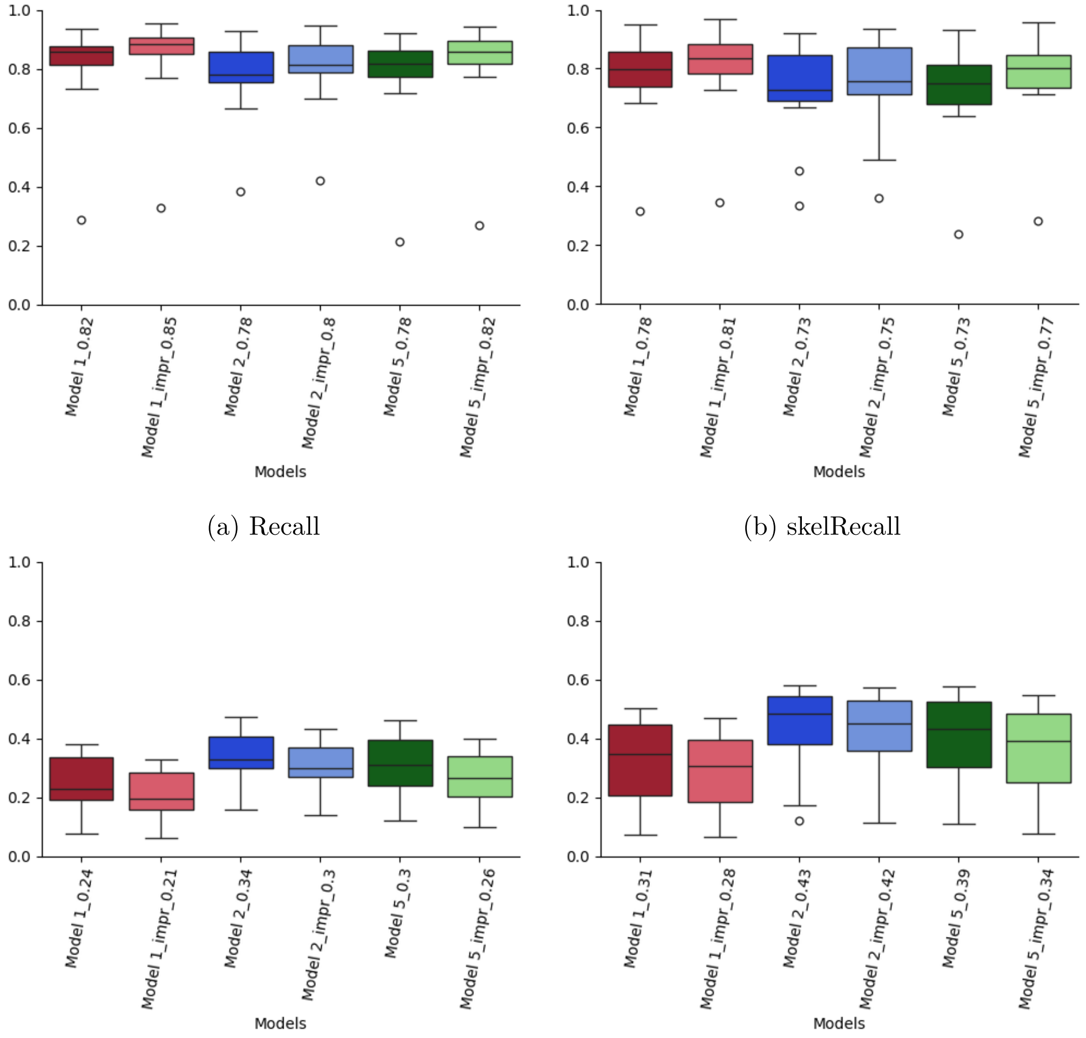
Table 9: Comparing the results of the clinical evaluation of the previous version of models 1, 2 and 5 with those where the threshold method is applied, the improved models (*1_impr*, *2_impr* and *5_impr*) are rated better.

Model Ensembles

In section 2.5.2, it is shown that the models with GIN augmentation (1 and 5) in some cases segment more parts of long thin vessels than the model with N3 applied (2), however, model 2 performs better when segmenting large vessels. Therefore, combining two or all of the three best models is considered to improve the result further. This is realized by joining the regions of the corresponding predictions.

The ensembles show better recall values compared to the original recall values as can be seen in the box plots in Figure 23. The largest recall value of the models before the improvement is achieved by model 1 with 0.82 and skelRecall value of 0.78. These values are exceeded for all model ensembles. At the same time, precision decreases only slightly in comparison to the corresponding original models.

In the clinical evaluation, the ensemble of model 1 and 2 is rated the best among all model combinations of two models. It is not surprising that the ensemble of all three models – Model 1+2+5 – yields the largest recall value combining the predictions of the other models. The exact evaluation results are shown in Table 10. As intended, the



(a) Recall

(b) skelRecall

(c) Precision

(d) skelPrecision

Figure 22: The application of the calibration method (_impr) with threshold 0.1 on the selected models 1 (7a), 2 (7b) and 5 (9a) shows an improvement of the recall values in comparison to the original models while the precision values decrease slightly. Thus, this method improves the models slightly with regard to the application in this project.

model ensembles improve the previous models, reducing underestimation. This becomes visible in Figure 24, where the segmentation results of the ensembles are compared to the singular models. The improved models segment in this case both large vessels as well as narrow structures almost completely.

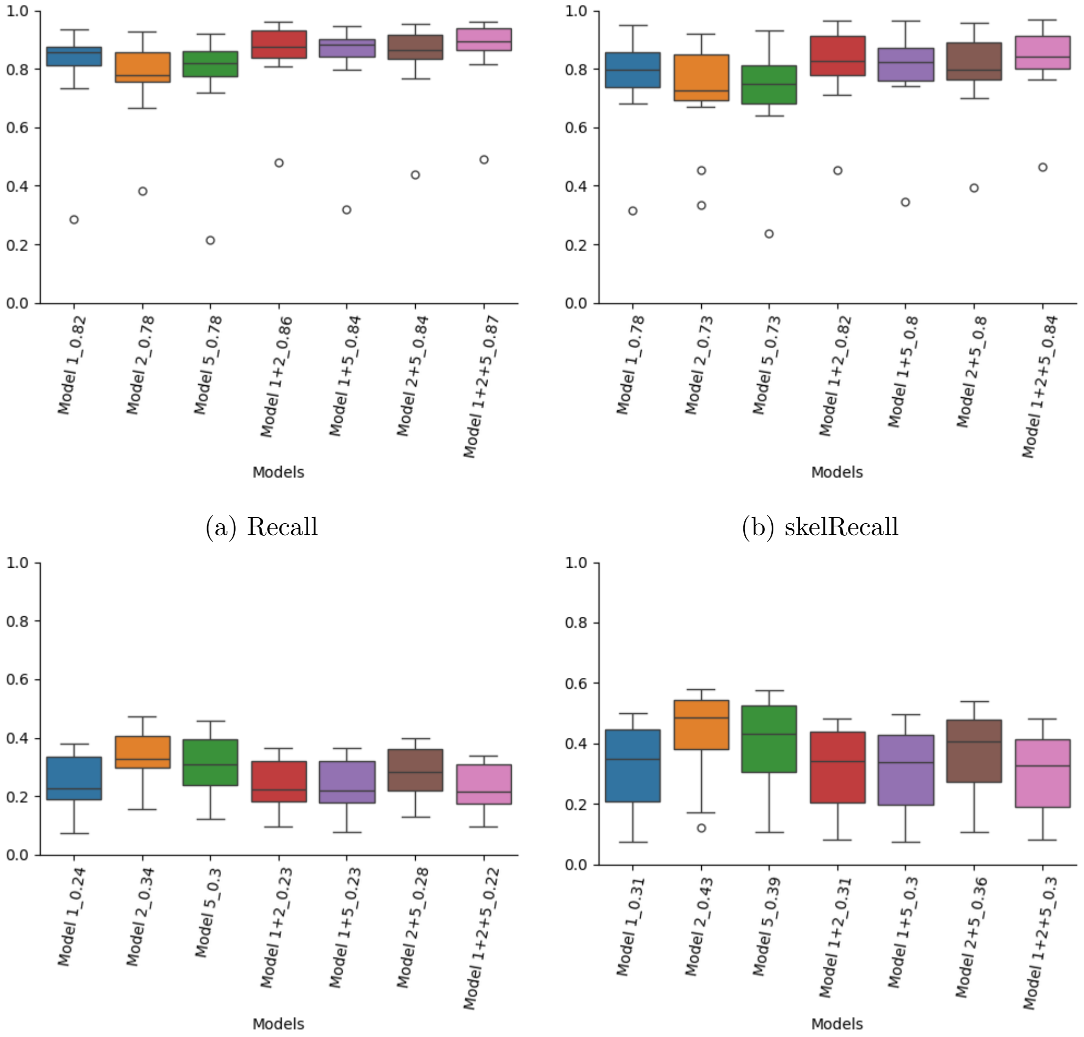
Model	**0**	**1**	**2**	**3**	**Total**
1+2	2	3	5	-	13
1+5	5	3	1	1	8
2+5	3	3	2	2	13
1+2+5	2	2	3	3	17

(a) Evaluation on PSC Data

Model	**0**	**1**	**2**	**3**	**Total**
1+2	-	2	8	-	18
1+5	1	3	6	-	15
2+5	1	2	7	-	16
1+2+5	-	2	8	-	18

(b) Evaluation on Control Group Data

Table 10: The combination of the best models 1,2 and 5 into model ensembles increases the rating scores in the clinical evaluation.



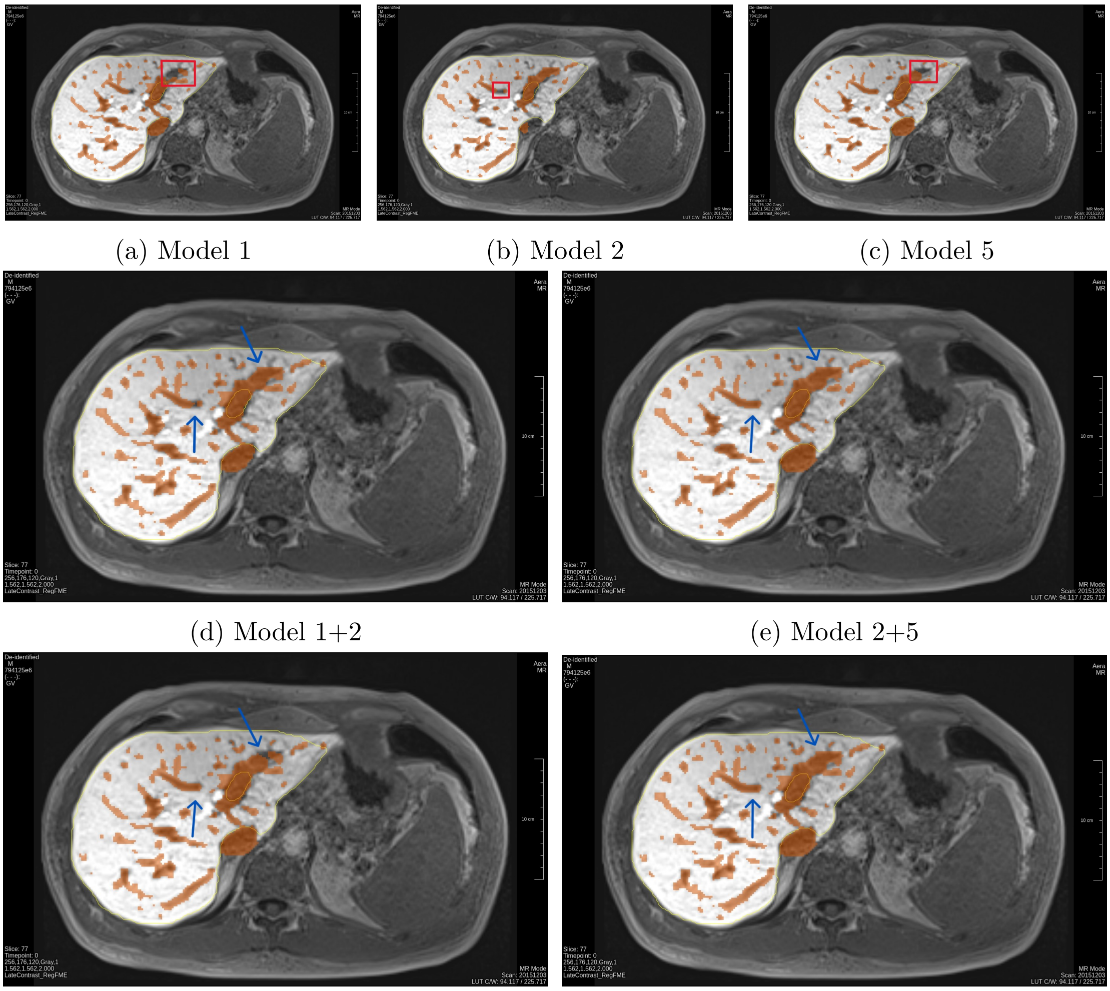
(a) Recall

(b) skelRecall

(c) Precision

(d) skelPrecision

Figure 23: The model ensembles increase the recall and skelRecall values in comparison to the original models while the precision and skelPrecision values decrease only slightly. This shows that the method of combining the models can visibly improve the results. Therefore, these models are preferred over the old models in the application of this project.



(a) Model 1

(b) Model 2

(c) Model 5

(d) Model 1+2

(e) Model 2+5

(f) Model 1+5

(g) Model 1+2+5

Figure 24: The model ensembles of the best models 1, 2 and 5 improve the vessel segmentation with regard to underestimation.

Chapter 3: Results: Liver Function Estimation and Impact of Vessel Segmentation

3.1 Automatic Workflow for Liver Function Estimation

The liver segmentation has been evaluated on the hepatobiliary phase first as the algorithm is developed for this. But due to problems with varying intensities on PSC data, it is run on the native phase (second rating) for the results presented in section 3.2. The segmentation on the native phase is rated better in the clinical evaluation. The rating scores are presented in Table 11.

Model	**0**	**1**	**2**	**3**
Liver	-	3	6	1
Liver Territories	10	-	-	-
Liver Lobes	1	-	3	6

(a) PSC Data (initial rating)

Model	**0**	**1**	**2**	**3**
Liver	-	1	3	6
Liver Territories	10	-	-	-
Liver Lobes	-	-	-	10

(b) Control Group Data (initial rating)

Model	**0**	**1**	**2**	**3**
Liver	-	1	2	7
Liver Sections	3	1	3	3
Liver Lobes	1	-	2	7

(c) PSC Data (second rating)

Model	**0**	**1**	**2**	**3**
Liver	-	1	1	8
Liver Sections	1	-	2	7
Liver Lobes	-	-	2	8

(d) Control Group Data (second rating)

Table 11: These rating scores of liver segmentation and the computation of liver sections and liver lobes are assigned by the medical student. Running the liver segmentation on the native phase, improves the results (c,d) compared to the initial rating phase (a,b).

The computation of the territories by Couinaud does not show good results (ref. to Tables 11a,11b) and is not usable within this application. Therefore, the division into four liver sections and two liver lobes is implemented as described in section 1.6. In this thesis, the liver function is evaluated globally on the whole liver segmentation but the good rating scores for liver sections and liver lobes create a good foundation for the computation of local liver function.

3.2 Automated vs. Manual Analysis of Liver Function

The segmentation methods described in section 3.1 as well as the vessel segmentation models developed in chapter 2 complete the automatic workflow enabling a user-independent determination of liver function values based on the dataset of PSC and control group patients. In this work, as liver function score, the Relative Enhancement (RE) is evaluated as defined in section 1.4. In the following, the results of the automatic workflow are analyzed in terms of their correlation with the manually determined ROI-based measurements.

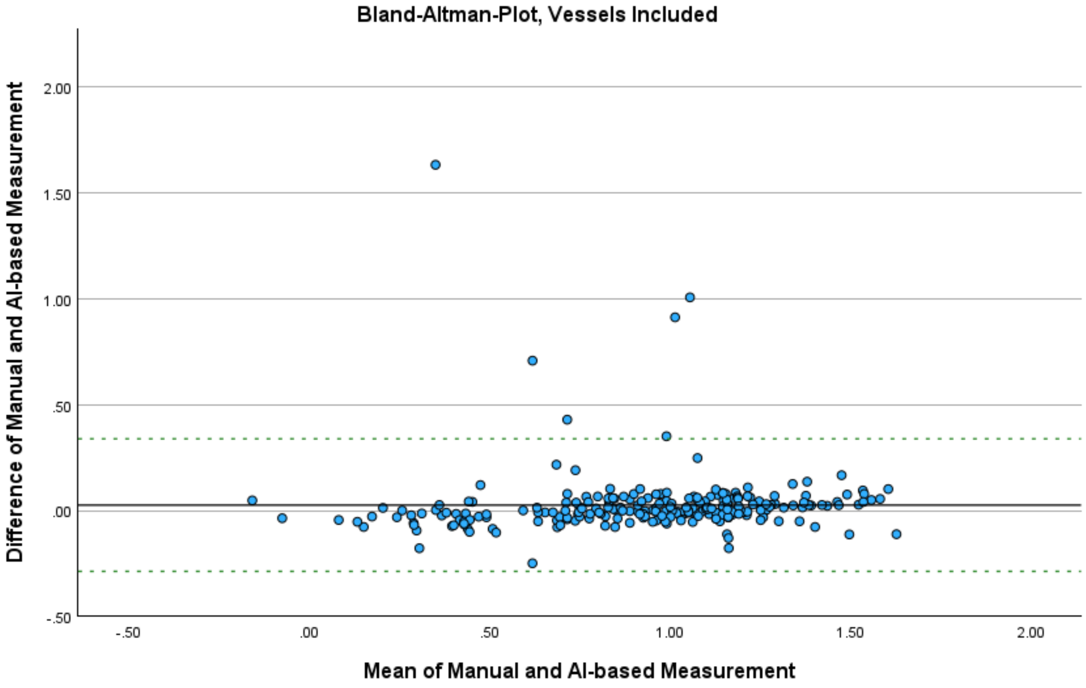


Figure 25: The Bland-Altman plot shows high similarity between the results of the automatic computation and the manual measurements of liver function (plot provided by Sina Dornbusch).

Comparing the RE values of the automatic and the manual approach, both methods show very similar results. This is visualized in a Bland-Altman plot in Figure 25. For this plot, five outliers have been removed because of very large (in the scale of 1000) or negative RE values in the automatic computation. These issues arise, for instance, during the registration process, leading to slices filled with zero values, and therefore resulting in invalid segmentations. The remaining data of 237 patients show only small differences between automated workflow and manual ROI-based measurements, with deviations of 0.0259 ± 0.026 in the mean. This data is provided by the medical student of this project.

The *intraclass coefficient* (ICC) assesses how similar the automatically and the manually determined RE values are among all cases [31]. The ICC values are computed by the medical student. As some outliers have been removed in the results of the automatic workflow, the One-Way Random-Effects model has been applied to compute the ICC values. This model can be used if not each subject is rated by all raters [31]. The ICC value for the automatic workflow is already very large with 0.946 and a 95% confidence interval of [0.930, 0.958] as presented in Table 12. So, even without consideration of a vessel segmentation model there is a high correlation between the liver function values of manual and automatic method.

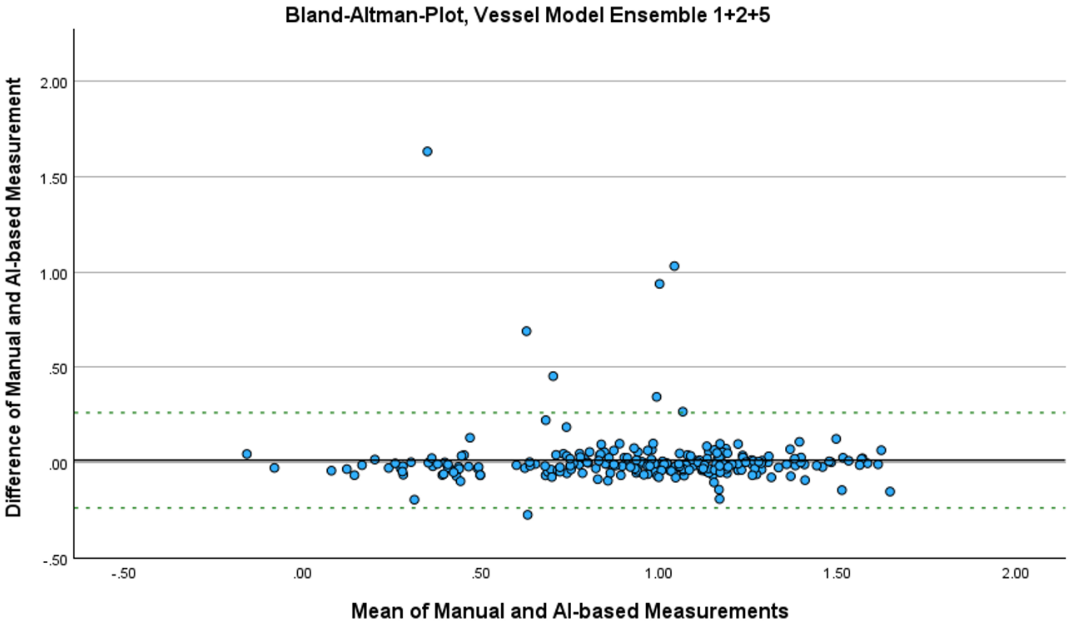


Figure 26: The Bland-Altman plot for the RE values with vessel segmentation model 1+2+5 applied shows similar results to the plot with vessels included (plot provided by Sina Dornbusch).

3.3 Influence of Vessel Segmentation on Liver Function Calculation

In the following, the impact of the different vessel segmentation models on the computation of liver function is analyzed. Considering the Bland-Altman-Plot in Figure 26 for the results with application of vessel segmentation model 1+2+5, it shows very similar differences between manual and AI-based measurements as without exclusion of vessels. The mean difference decreases a bit to 0.0125 ± 0.26 in comparison to the mean difference of 0.0259 ± 0.026 with vessels included, indicating a slight improvement by applying the vessel segmentation model.

The ICC also changes only slightly for application of different vessel segmentation models. Contrary to expectations, the model calibration by a threshold method in models *1_impr*, *2_impr* and *5_impr* decreases the ICC slightly despite better rating scores than the previous models 1, 2 and 5. This can be seen in comparing the values in Table 12. This decline is probably due to the fact that the improved models segment not only more vessels but also more liver tissue especially in the marginal area of the vessels leading to variations in the ICC values. However, the differences are marginal and the improved models still show similar or better ICC values than for the computation of liver function without exclusion of vessels.

The model ensembles show also very similar results as displayed in Table 12. All ICC values are above 0.9 and therefore indicate very good correlation results.

Model		**ICC**	**95% CI**
Vessels Included	Single Measures	0.897	[0.869, 0.919]
	Average Measures	0.946	[0.930, 0.958]
1	Single Measures	0.900	[0.873, 0.922]
	Average Measures	0.948	[0.932, 0.959]
1_impr	Single Measures	0.897	[0.869, 0.919]
	Average Measures	0.946	[0.930, 0.958]
2	Single Measures	0.900	[0.873, 0.922]
	Average Measures	0.948	[0.932, 0.959]
2_impr	Single Measures	0.897	[0.869, 0.920]
	Average Measures	0.946	[0.930, 0.958]
5	Single Measures	0.901	[0.874, 0.922]
	Average Measures	0.948	[0.933, 0.960]
5_impr	Single Measures	0.898	[0.870, 0.920]
	Average Measures	0.946	[0.931, 0.958]
1+2	Single Measures	0.900	[0.873, 0.922]
	Average Measures	0.948	[0.932, 0.959]
1+5	Single Measures	0.900	[0.873, 0.922]
	Average Measures	0.948	[0.932, 0.959]
2+5	Single Measures	0.901	[0.874, 0.922]
	Average Measures	0.948	[0.932, 0.960]
1+2+5	Single Measures	0.900	[0.873, 0.922]
	Average Measures	0.948	[0.932, 0.959]

Table 12: There is a high correlation between the liver function results of the automatic workflow and the manual measurements. The improved models *1_impr*, *2_impr* and *5_impr* show marginally lower ICC values than models 1, 2 and 5 and the model ensembles (data provided by Sina Dornbusch).

Model	**Liver Tissue** (mm^3)	**Vessels** (mm^3)
Vessels Included	1,663,230.1	-
1	1,449,862.6	213,367.5
2	1,459,384.0	203,846.1
5	1,493,849.8	169,380.4
1_impr	1,306,832.2	356,397.9
2_impr	1,334,449.0	328,781.1
5_impr	1,369,043.1	294,187.0
1+2	1,399,929.1	263,301.1
1+5	1,428,736.5	234,493.7
2+5	1,429,689.7	233,540.4
1+2+5	1,389,316.3	273,913.9

Table 13: Mean volume of the liver tissue and the vessels over all patients in the PSC and control group dataset showing that the vessel masks of the different models make up a notable part of the liver mask.

The low impact of excluding the vessels from the liver mask on the estimation of liver function is surprising, as the image intensities of the liver tissue differ considerably from those of the vessels. Comparing the mean volume of the excluded vessels to the liver volume in Table 13 shows that a relevant amount of voxels is segmented as vessels which accounts for approximately 12-20% of the original liver mask. This would also suggest a larger effect of the vessel mask on the correlation results. To explain the similar correlation results, a clinical evaluation of the entire dataset instead of only 20 patients could be helpful and provide more detailed information on possible reasons. Furthermore, lesion segmentation is not considered in these results, yet, which is discussed in more detail in the following chapter.

Chapter 4: Conclusion and Discussion

The results of the automatic workflow for computation of liver function presented in this thesis show strong correlation with the manual ROI-based measurements. This is a solid basis for future applications of this kind of assessment of the liver function in clinical practice. However, for medical application, the correlation with clinical endpoints would be interesting.

The vessel segmentation models developed within this work improve the correlation results slightly as explained in chapter 3. The model ensemble of models 1, 2 and 5 shows the highest recall values on the test dataset, the highest rating scores in the clinical evaluation by the medical student and strong correlation results. Nevertheless, computing the segmentation results of three models increases the computation time significantly, taking three times as long as computing the results of a singular model. Therefore, it could be interesting to use the best singular model – model *1_impr*. This shows equivalent correlation results within this work and is less time consuming.

As prospect, it would be interesting to train model 1 again with cases from the PSC dataset as additional training data. For this, the segmentation results of the existing model on PSC data could be corrected and these annotations could be used as new reference. In this way, the model could learn to avoid the errors made before. The new annotations could be done as correction for under- and overestimation and would probably improve recall and precision at the same time. This could improve the Precision-Recall curves and especially the AP value described in section 2.6. However, as the segmentation models presented in this work did not show a high impact on the resulting liver function values in the workflow, it is questionable whether further improvement of the models can have more influence.

Another remaining issue is the impact of lesion segmentation on the liver function results. As described in section 1.6, for the lesion segmentation, a seed point needs to be set manually in each lesion structure. This is very time consuming and could not be taken into account within this thesis. In the long-term, a fully automatic segmentation could simplify the workflow. However, since the vessel segmentation has low impact on the correlation results, the impact of a lesion segmentation might similarly be lower than expected. A higher impact of both – vessels and lesions – could appear in the evaluation of local liver function. This is done as part of the work of the medical student as this requires more detailed rating results of all cases to evaluate which liver territories or sections are suitable for accurate results.

Although there are still some ideas for advancing the models and improving the results, several goals were achieved in this work. Various vessel segmentation models have been developed and improved with the best models reaching large recall values and good rating scores. Furthermore, high correlation results of the automatic workflow to the manual ROI-based analysis have been shown. All in all, the results show that this approach of automatic computation of liver function is promising and worth to be pursued further.

References

[1] Abraham, N. and Khan, N.M. (2019). A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation. In 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), pp. 683–687. DOI: 10.1109/ISBI.2019.8759329.

[2] Bartulos, C., Senk, K., Schumacher, M., Plath, J., Kaiser, N., Bade, R., Woetzel, J., and Wiggermann, P. (2022). Assessment of Liver Function With MRI: Where Do We Stand? Frontiers in Medicine *9*, 839919. DOI: 10.3389/fmed.2022.839919.

[3] Chazouilleres, O., Beuers, U., Bergquist, A., Karlsen, T.H., Levy, C., Samyn, M., Schramm, C., and Trauner, M. (2022). EASL Clinical Practice Guidelines on Sclerosing Cholangitis. Journal of Hepatology *77*, 761–806. ISSN: 0168-8278. DOI: https://doi.org/10.1016/j.jhep.2022.05.011.

[4] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A.L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence *40*, 834–848. DOI: 10.1109/TPAMI.2017.2699184.

[5] Chierici, A., Lareyre, F., Salucki, B., Iannelli, A., Delingette, H., and Raffort, J. (2024). Vascular liver segmentation: a narrative review on methods and new insights brought by artificial intelligence. Journal of International Medical Research *52*, 1–23. DOI: 10.1177/03000605241263170.

[6] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, pp. 424–432. DOI: 10.1007/978-3-319-46723-8_49.

[7] Ciecholewski, M. and Kassjański, M. (2021). Computational Methods for Liver Vessel Segmentation in Medical Imaging: A Review. Sensors *21*. ISSN: 1424-8220. DOI: 10.3390/s21062027. URL: https://www.mdpi.com/1424-8220/21/6/2027.

[8] Couinaud, C. (1957). Le foie: études anatomiques et chirurgicales. Masson.

[9] Dice, L.R. (1945). Measures of the Amount of Ecologic Association Between Species. Ecology *26*, 297–302. ISSN: 00129658, 19399170. DOI: 10.2307/1932409.

[10] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters *27*. ROC Analysis in Pattern Recognition, 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010.

[11] Fraz, M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A., Owen, C., and Barman, S. (2012). Blood vessel segmentation methodologies in retinal images – A survey. Computer Methods and Programs in Biomedicine *108*, 407–433. ISSN: 0169-2607. DOI: `https://doi.org/10.1016/j.cmpb.2012.03.009`.

[12] Ginsburg, B., Castonguay, P., Hrinchuk, O., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., Nguyen, H., Zhang, Y., and Cohen, J.M. (2020). *Training Deep Networks with Stochastic Gradient Normalized by Layerwise Adaptive Second Moments.* URL: `https://openreview.net/forum?id=BJepq2VtDB`.

[13] Goceri, E., Shah, Z.K., and Gurcan, M.N. (2017). Vessel segmentation from abdominal magnetic resonance images: adaptive and reconstructive approach. International journal for numerical methods in biomedical engineering *33*, e2811. DOI: `https://doi.org/10.1002/cnm.2811`.

[14] Guyton, A.C. and Hall, J.E. (2006). Text book of medical physiology. Philadelphia: Elsevier Saunders, pp. 759–761, 837–842.

[15] Hänsch, A., Thielke, F., Meine, H., Rennebaum, S., Froelich, M.F., Becker, L.S., Hinrichs, J.B., and Schenk, A. (2022). Robust Liver Segmentation with Deep Learning Across DCE-MRI Contrast Phases. In Bildverarbeitung für die Medizin 2022, (K. Maier-Hein, T.M. Deserno, H. Handels, A. Maier, C. Palm, and T. Tolxdorff, eds.). Wiesbaden: Springer Fachmedien Wiesbaden, pp. 13–18. DOI: `10.1007/978-3-658-36932-3_3`.

[16] Hering, A., Peisen, F., Amaral, T., Gatidis, S., Eigentler, T., Othman, A., and Moltz, J.H. (2021). Whole-Body Soft-Tissue Lesion Tracking and Segmentation in Longitudinal CT Imaging Studies. In Proceedings of the Fourth Conference on Medical Imaging with Deep Learning, (M. Heinrich, Q. Dou, M. de Bruijne, J. Lellmann, A. Schläfer, and F. Ernst, eds.). Vol. 143. Proceedings of Machine Learning Research. PMLR, pp. 312–326. URL: `https://proceedings.mlr.press/v143/hering21a.html`.

[17] Hering, A., Westphal, M., Gerken, A., Almansour, H., Maurer, M., Geisler, B., Kohlbrandt, T., Eigentler, T., Amaral, T., Lessmann, N., et al. (2024). Improving assessment of lesions in longitudinal CT scans: A bi-institutional reader study on an AI-assisted registration and volumetric segmentation workflow. International Journal of Computer Assisted Radiology and Surgery *19*, 1689–1697. DOI: `https://doi.org/10.1007/s11548-024-03181-4`.

[18] Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, (F. Bach and D. Blei, eds.). Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 448–456. URL: `https://proceedings.mlr.press/v37/ioffe15.html`.

[19] Isensee, F., Jaeger, P., Kohl, S., Petersen, J., and Maier-Hein, K. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods *18*, 1–9. DOI: 10.1038/s41592-020-01008-z.

[20] Isensee, F., Jäger, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., Schock, J., Klein, A., Roß, T., Wirkert, S., et al. (2020). *Batchgenerators - a Python Framework for Data Augmentation.* Version 0.19.6. DOI: 10.5281/zenodo.3632567.

[21] Isensee, F., Petersen, J., Kohl, S., Jaeger, P., and Maier-Hein, K. (2019). nnU-Net: Breaking the Spell on Successful Medical Image Segmentation. Nature Methods *18*, 203–211. DOI: 10.48550/arXiv.1904.08128.

[22] Jadon, S. (2020). A survey of loss functions for semantic segmentation. In 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, pp. 1–7. DOI: 10.1109/cibcb48159.2020.9277638.

[23] Kamath, P.S., Wiesner, R.H., Malinchoc, M., Kremers, W., Therneau, T.M., Kosberg, C.L., D'Amico, G., Dickson, E., and Kim, W. (2001). A Model to Predict Survival in Patients with End-Stage Liver Disease. Hepatology *33*, 464–470. DOI: https://doi.org/10.1053/jhep.2001.22172.

[24] Kazami, Y., Kaneko, J., Keshwani, D., Takahashi, R., Kawaguchi, Y., Ichida, A., Ishizawa, T., Akamatsu, N., Arita, J., and Hasegawa, K. (2021). Artificial intelligence enhances the accuracy of portal and hepatic vein extraction in computed tomography for virtual hepatectomy. Journal of Hepato-Biliary-Pancreatic Sciences *29*, 359–368. DOI: 10.1002/jhbp.1080.

[25] Kim, W.R., Therneau, T.M., Wiesner, R.H., Poterucha, J.J., Benson, J.T., Malinchoc, M., Larusso, N.F., Lindor, K.D., and Dickson, E.R. (2000). A Revised Natural History Model for Primary Sclerosing Cholangitis. Mayo Clinic Proceedings *75*, 688–694. ISSN: 0025-6196. DOI: https://doi.org/10.4065/75.7.688.

[26] Kingma, D.P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), DOI: https://doi.org/10.48550/arxiv.1412.6980.

[27] Kirchhoff, Y., Rokuss, M.R., Roy, S., Kovacs, B., Ulrich, C., Wald, T., Zenk, M., Vollmuth, P., Kleesiek, J., Isensee, F., et al. (2024). Skeleton Recall Loss for Connectivity Conserving and Resource Efficient Segmentation of Thin Tubular Structures. In Computer Vision – ECCV 2024, (A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, eds.). Cham: Springer Nature Switzerland, pp. 218–234. DOI: https://doi.org/10.1007/978-3-031-72980-5_13.

[28] Klein, J., Wenzel, M., Romberg, D., Köhn, A., Kohlmann, P., Link, F., Hänsch, A., Dicken, V., Stein, R., Haase, J., et al. (2020). QuantMed: Component-based Deep Learning Platform for Translational Research. In Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, (P.-H. Chen and T.M. Deserno, eds.). Vol. 11318. International Society for Optics and Photonics. SPIE, pp. 229–236. DOI: `10.1117/12.2549582`.

[29] Kock, F., Thielke, F., Abolmaali, N., Meine, H., and Schenk, A. (2024). Suitability of DNN-based vessel segmentation for SIRT planning. International journal of computer assisted radiology and surgery *19*, 233–240. DOI: `https://doi.org/10.1007/s11548-023-03005-x`.

[30] Kock, F., Thielke, F., Chlebus, G., and Meine, H. (2022). Confidence Histograms for Model Reliability Analysis and Temperature Calibration. In Proceedings of The 5th International Conference on Medical Imaging with Deep Learning, (E. Konukoglu, B. Menze, A. Venkataraman, C. Baumgartner, Q. Dou, and S. Albarqouni, eds.). Vol. 172. Proceedings of Machine Learning Research. PMLR, pp. 741–759. URL: `https://proceedings.mlr.press/v172/kock22a.html`.

[31] Koo, T.K. and Li, M.Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine *15*, 155–163. ISSN: 1556-3707. DOI: `https://doi.org/10.1016/j.jcm.2016.02.012`.

[32] Lebre, M.-A., Vacavant, A., Grand-Brochier, M., Rositi, H., Abergel, A., Chabrot, P., and Magnin, B. (2019). Automatic segmentation methods for liver and hepatic vessels from CT and MRI volumes, applied to the Couinaud scheme. Computers in Biology and Medicine *110*, 42–51. ISSN: 0010-4825. DOI: `https://doi.org/10.1016/j.compbiomed.2019.04.014`.

[33] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. (July 2018). Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence *PP*, 318–327. DOI: `10.1109/TPAMI.2018.2858826`.

[34] Liu, W., Rabinovich, A., and Berg, A.C. (2015). *Parsenet: Looking wider to see better.* URL: `https://arxiv.org/abs/1506.04579`.

[35] Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., and Martel, A.L. (2021). Loss Odyssey in Medical Image Segmentation. Medical Image Analysis *71*, 102035. ISSN: 1361-8415. DOI: `https://doi.org/10.1016/j.media.2021.102035`.

[36] Maier-Hein, L., Reinke, A., Godau, P., et al. (2024). Metrics Reloaded: Recommendations for Image Analysis Validation. Nature Methods *21*, 195–212. DOI: `10.1038/s41592-023-02151-z`.

[37] Manns, M.P., Bergquist, A., Karlsen, T.H., Levy, C., Muir, A.J., Ponsioen, C., Trauner, M., Wong, G., and Younossi, Z.M. (2025). Primary Sclerosing Cholangitis. Nature Reviews Disease Primers *11*, 17. DOI: `https://doi.org/10.1038/s41572-025-00600-x`.

[38] MeVis Medical Solutions AG and Fraunhofer MEVIS (2003–2025). *MeVisLab – Medical Imaging and Visualization Software*. `https://www.mevislab.de`. Version 4.1.70.411 (2025-03-30 Release).

[39] Mishra, P. and Sarawadekar, K. (2019). Polynomial learning rate policy with warm restart for deep neural network. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON), IEEE, pp. 2087–2092. DOI: `10.1109/TENCON.2019.8929465`.

[40] Moccia, S., De Momi, E., El Hadji, S., and Mattos, L.S. (2018). Blood vessel segmentation algorithms — Review of methods, datasets and evaluation metrics. Computer Methods and Programs in Biomedicine *158*, 71–91. ISSN: 0169-2607. DOI: `https://doi.org/10.1016/j.cmpb.2018.02.001`.

[41] Moraes, T., Amorim, P., Da Silva, J.V., and Pedrini, H. (2020). Medical image interpolation based on 3D Lanczos filtering. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization *8*, 294–300. DOI: `https://doi.org/10.1080/21681163.2019.1683469`.

[42] Nazir, A., Cheema, M.N., Sheng, B., Li, P., Kim, J., and Lee, T.-Y. (2021). Living Donor-Recipient Pair Matching for Liver Transplant via Ternary Tree Representation With Cascade Incremental Learning. IEEE Transactions on Biomedical Engineering *68*, 2540–2551. DOI: `10.1109/TBME.2021.3050310`.

[43] Oh, N., Kim, J.-H., Rhu, J., Jeong, W.K., Choi, G.-s., Kim, J.M., and Joh, J.-W. (2023). Automated 3D liver segmentation from hepatobiliary phase MRI for enhanced preoperative planning. Scientific Reports *13*, 17605. DOI: `10.1038/s41598-023-44736-w`.

[44] Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., and Rueckert, D. (2022). Causality-inspired Single-source Domain Generalization for Medical Image Segmentation. IEEE Transactions on Medical Imaging *42*, 1095–1106. DOI: `https://doi.org/10.1109/TMI.2022.3224067`.

[45] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, (N. Navab, J. Hornegger, W.M. Wells, and A.F. Frangi, eds.). Cham: Springer International Publishing, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: `https://doi.org/10.1007/978-3-319-24574-4_28`.

[46] Salehi, S.S.M., Erdogmus, D., and Gholipour, A. (2017). Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. (Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, eds.), 379–387. DOI: https://doi.org/10.1007/978-3-319-67389-9_44.

[47] Sander, J., Vos, B.D. de, Wolterink, J.M., and Išgum, I. (2019). Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. In Medical Imaging 2019: Image Processing, (E.D. Angelini and B.A. Landman, eds.). Vol. 10949. International Society for Optics and Photonics. SPIE, p. 1094919. DOI: 10.1117/12.2511699.

[48] Schulze, J., Lenzen, H., Hinrichs, J.B., Ringe, B., Manns, M.P., Wacker, F., and Ringe, K.I. (2019). An Imaging Biomarker for Assessing Hepatic Function in Patients with Primary Sclerosing Cholangitis. Clinical Gastroenterology and Hepatology *17*, 192–199. DOI: https://doi.org/10.1016/j.cgh.2018.05.011.

[49] Selle, D., Preim, B., Schenk, A., and Peitgen, H.-O. (2002). Analysis of vasculature for liver surgical planning. IEEE Transactions on Medical Imaging *21*, 1344–1357. DOI: 10.1109/TMI.2002.801166.

[50] Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P.W., Bauer, U., and Menze, B.H. (2021). clDice - A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16560–16569. DOI: https://doi.org/10.1109/CVPR46437.2021.01629.

[51] Singh, B., De, S., Zhang, Y., Goldstein, T., and Taylor, G. (2015). Layer-Specific Adaptive Learning Rates for Deep Networks. In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 364–368. DOI: 10.1109/ICMLA.2015.113.

[52] Sled, J., Zijdenbos, A., and Evans, A. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging *17*, 87–97. DOI: 10.1109/42.668698.

[53] Soler, L., Nicolau, S., Pessaux, P., Mutter, D., and Marescaux, J. (2014). Real-time 3D image reconstruction guidance in liver resection surgery. Hepatobiliary Surgery and Nutrition *3*, 73–81. ISSN: 2304-389X. DOI: 10.3978/j.issn.2304-3881.2014.02.03.

[54] Spahr, N., Thoduka, S., Abolmaali, N., Kikinis, R., and Schenk, A. (2019). Multimodal image registration for liver radioembolization planning and patient assessment. International journal of computer assisted radiology and surgery *14*, 215–225. DOI: https://doi.org/10.1007/s11548-018-1877-5.

[55] Strasberg, S., Belghiti, J., Clavien, P.-A., Gadzijev, E., Garden, J., Lau, W.-Y., Makuuchi, M., and Strong, R. (2000). The Brisbane 2000 Terminology of Liver Anatomy and Resections. HPB *2*, 333–339. ISSN: 1365-182X. DOI: `https://doi.org/10.1016/S1365-182X(17)30755-4`.

[56] Strehlow, J., Spahr, N., Rühaak, J., Laue, H., Abolmaali, N., Preusser, T., and Schenk, A. (2018). Landmark-based evaluation of a deformable motion correction for DCE-MRI of the liver. International journal of computer assisted radiology and surgery *13*, 597–606. DOI: `https://doi.org/10.1007/s11548-018-1710-1`.

[57] Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. (2017). Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, pp. 240–248. DOI: `https://doi.org/10.1007/978-3-319-67558-9_28`.

[58] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In Proceedings of the 30th International Conference on Machine Learning, (S. Dasgupta and D. McAllester, eds.). Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, pp. 1139–1147. URL: `https://proceedings.mlr.press/v28/sutskever13.html`.

[59] Taghanaki, S.A., Zheng, Y., Kevin Zhou, S., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., and Hamarneh, G. (2019). Combo Loss: Handling Input and Output Imbalance in Multi-organ Segmentation. Computerized Medical Imaging and Graphics *75*, 24–33. ISSN: 0895-6111. DOI: `https://doi.org/10.1016/j.compmedimag.2019.04.005`.

[60] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C. (2015). Efficient Object Localization Using Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 648–656.

[61] Wittmann, B., Wattenberg, Y., Amiranashvili, T., Shit, S., and Menze, B. (2025). vesselFM: A Foundation Model for Universal 3D Blood Vessel Segmentation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 20874–20884. DOI: `https://doi.org/10.48550/arXiv.2411.17386`.

[62] Yeung, M., Sala, E., Schönlieb, C.-B., and Rundo, L. (2022). Unified Focal Loss: Generalising Dice and Cross Entropy-based Losses to Handle Class Imbalanced Medical Image Segmentation. Computerized Medical Imaging and Graphics *95*, 102026. ISSN: 0895-6111. DOI: `https://doi.org/10.1016/j.compmedimag.2021.102026`.

[63] Yu, A.W., Huang, L., Lin, Q., Salakhutdinov, R., and Carbonell, J. (2018). Block-Normalized Gradient Method: An Empirical Study for Training Deep Neural Network. arXiv: 1707.04822 [cs.LG]. URL: https://arxiv.org/abs/1707.04822.

[64] Zbinden, L., Catucci, D., Suter, Y., Berzigotti, A., Ebner, L., Christe, A., Obmann, V., Sznitman, R., and Huber, A. (2022). Convolutional neural network for automated segmentation of the liver and its vessels on non-contrast T1 vibe Dixon acquisitions. Scientific Reports *12*, 22059. DOI: 10.1038/s41598-022-26328-2.

[65] Zhao, Z., Li, W., Ding, X., Sun, J., and Xu, L.X. (2025). TTGA U-Net: Two-stage two-stream graph attention U-Net for hepatic vessel connectivity enhancement. Computerized Medical Imaging and Graphics *122*, 102514. ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.compmedimag.2025.102514.

[66] Zhu, M. (2004). Recall, Precision and Average Precision. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo *2*, 6. URL: https://datascience-intro.github.io/1MS041-2022/Files/AveragePrecision.pdf.