Development of AI-based methods for processing and quantitative analysis of radiological image data

Doctoral thesis

to obtain a doctorate (PhD)

from the Faculty of Medicine

of the University of Bonn

Sebastian Nowak

from Neuwied, Germany

2023

Written with authorization of
the Faculty of Medicine of the University of Bonn

First reviewer: Prof. Dr. med. Ulrike I. Attenberger

Second reviewer: Prof. Dr. rer. nat. Jürgen Hesser

Day of oral examination: 15.03.2023

From the Department of Diagnostic and Interventional Radiology,
University Hospital Bonn
Director: Univ.-Prof. Dr. med. Ulrike I. Attenberger

Table of Contents

List of abbreviations

QIA	Quantitative Image Analysis
AI	Artificial Intelligence
DL	Deep Learning
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
CNN	Convolutional Neural Network
CDFNet	Competitive Dense Fully Connected Network
FMF	Fatty Muscle Fraction
SAT	Subcutaneous Adipose Tissue
SM	Skeletal Muscle
VAT	Visceral Adipose Tissue
ACC	Accuracy
AP	Average Precision
AUC	Area Under the Curve
DTL	Deep Transfer Learning
MRL	Magnetic Resonance Lymphangiography
SCT	Subcutaneous Tissue
SFT	Subfascial Tissue
ANI	Artificial Narrow Intelligence
AGI	Artificial General Intelligence

1. Abstract

Quantitative Image Analysis (QIA), that is, software-based extraction and analysis of numerically quantifiable features from medical imaging, has great potential to contribute to the progression of precision and personalized medicine. By utilizing objective, quantifiable features, biomarkers can be defined or predictive models can be developed that allow, e.g., the automatic detection of pathological alterations or the monitoring of disease progression or therapeutic success. To be practical in routine clinical care, QIA of tissues and organs should be automated and require minimal intervention by the radiologist. Artificial Intelligence (AI) methods and Deep Learning (DL) in particular have emerged as state-of-the-art image processing techniques in recent years, also for medical imaging.

This work features three AI-based pipelines for automated QIA. First, an end-to-end automated pipeline for quantification of muscle and adipose tissue (termed body composition analysis) is presented that includes automatic 2D slice extraction from 3D Computed Tomography (CT) scans, automatic tissue segmentation and quality control mechanisms to warn of potential invalid analysis. Then, a DL pipeline for automatic detection of liver cirrhosis in Magnetic Resonance Imaging (MRI) is demonstrated that features a method of explainable AI proposed to highlight image regions of importance. Finally, a pipeline for quantitative tissue assessment in MRI allowing also for monitoring of therapeutic success in patients with lip- and lymphedema is developed. This work includes a two-step anatomical landmark detection in combination with quality-assured tissue segmentation to create visualizations of tissue distribution in a standardized leg model.

The results of this work provide insights into the development of automated AI-based pipelines for use in clinical routine. Besides investigating methods for tissue segmentation, anatomical landmark and disease detection, it was also explored how combinations of those methods can be used in pipelines that overcome challenges of routine clinical data and thus minimize required effort of the radiologist as a prerequisite for potential use in clinical practice.

2. Introduction and aims with references

Conventionally, radiologists evaluate tissues within medical images primarily visually (Hosny et al., 2018). Precise numerical quantifications of tissue properties by the attending physician beyond simple diameter or region of interest measurements are typically labour intensive and thus not feasible in clinical practice. QIA aims to extract quantifiable information from radiological images through automated software-based analysis. The utilization of otherwise unused quantifiable features has the potential to contribute to a more objective evaluation of pathologic alterations. Moreover, by retrospective investigation of the correlation of disease progression with quantitative features, image-based biomarkers can be identified, which may then be assessed as useful indicators for diagnostic or treatment decisions of future patients. Thus, QIA methods represent important tools in the field of precision and personalized medicine (Hagiwara et al., 2020).

A prominent example is the body composition analysis, in which connective tissue compartments are quantitatively assessed in abdominal cross-sectional imaging. The amount and quality of adipose and muscle tissue has shown to have prognostic implications with regard to various oncologic and cardiovascular diseases (Faron et al., 2021; Luetkens et al., 2020; Prado et al., 2008). Furthermore, in combination with AI tools, quantifiable features can be used to develop automated decision systems that can detect a disease in imaging, characterize the disease, such as staging or etiology identification, or monitor disease progression or therapy success (Hosny et al., 2018).

The first step of a QIA pipeline is often the identification of the tissue, organ or lesion of interest by segmentation. The clinical usability of quantitative methods is drastically limited if time intensive manual segmentations are required. Thus, automatic segmentation is an essential pre-requisite for assessment of quantitative biomarkers in clinical routine. In the past, automatic medical image analyses were realized with sequential application of simple image processing algorithms, such as edge and line detector filters and mathematical modelling (Litjens et al., 2017). During the rise of AI applications in recent years, DL methods and especially Convolutional Neural Networks (CNN) have emerged as state-of-the-art image

segmentation and general image processing techniques. However, to train a CNN for image segmentation by supervised learning, a large number of already segmented images are required. For annotation of non-medical images, e.g. traffic scenes, extensive expert knowledge is usually not necessary, allowing large data sets to be compiled, e.g. in crowdsourcing approaches. However, the assessment of medical images usually requires significant domain knowledge (Hosny et al., 2018). Moreover, medical images are usually subject to strict privacy policies making the compilation of large datasets from different international centers difficult. As a result, AI methods developed on and for images of clinical routine often have to be trained on-site in clinics and with costly annotation by physicians. Efficient use of annotation time is therefore of particular importance (see Figure 2.1).

Based on the segmentation of the tissue of interest, methods of QIA can be applied to investigate the diagnostic or prognostic value of user-defined 'constructed' image features, such as tissue volumes, densities or intensity histogram analyses. For example, the prognostic value of individual constructed quantitative features can be analysed using conventional statistical methods such as Kaplan-Meier analyses in outcome studies (Luetkens et al., 2020). A technique termed Radiomics represents a more comprehensive approach to examine user-defined constructed features for the development of predictive models. Here, a vast variety of standardized quantitative features such as intensity histograms, textures, and shapes are collectively analysed using Machine Learning. Radiomics is premised on the hypothesis that medical images are multidimensional data that may contain features with correlations to pathophysiologies, and aims to uncover these correlations through data-mining of the digital image values (Gillies et al., 2015).

Another way to utilize relevant features from multidimensional data is the use of DL. The application of DL and especially CNNs fundamentally differs from the previous mentioned analysis of user-defined constructed features. Here, the method itself learns to identify and recognize relevant features within the image by optimizing its trainable parameters. This approach is capable of analysing highly abstract feature representations by combining self-learned features in deep layers.

DL is considered to have greater potential to create decision systems with clinical utility compared to traditional analysis of constructed human-defined features (Hosny et al., 2018). However, the processing of self-learned abstract feature representations result in the interpretation of the rationale behind the decision not being straightforward for humans, which is why DL methods are often referred to as "black box" (Petch et al., 2022). Improving the explainability of the models prediction, by methods which, for example, highlight areas of the image that seemed relevant, are therefore important aspects for DL-based QIA.

There are other challenges that emerge when using routine clinical data. One is that imaging is not standardized, but that e.g. the resolution and scan lengths can be variable, resulting in variable image sizes. Another challenge is that they may feature artefacts that make analysis of a tissue difficult or even impossible. Systems that operate with minimal interaction of the radiologist should provide solutions to these challenges. Thus, quality control mechanisms that detect and warn of potentially invalid analyses are an important part of QIA pipelines with utility for clinical routine.

Therefore, the aim of this thesis was to develop concepts and provide insights into the development of AI-based methods for the analysis and processing of routine clinical radiological images. Besides investigating methods for tissue segmentation, anatomical landmark and disease detection, this work also explores how analysis pipelines can be created in combination with other methods that overcome the challenges of clinical data and thus potentially be beneficial for clinical practice. The following systems are presented in this thesis:

i. An end-to-end automated pipeline for body composition analysis featuring automated 2D slice extraction from a 3D CT scan, adipose and muscle tissue segmentation and quality control mechanisms.
ii. A pipeline for the detection of liver cirrhosis in MRI including transfer learning and a method of explainable AI.
iii. A pipeline for standardized assessment and visualization of leg tissue distribution in patients with lip- and lymphedema in MRI featuring a two-step anatomical landmark detection and tissue segmentation with quality control that can be used to monitor disease progression and therapy success.

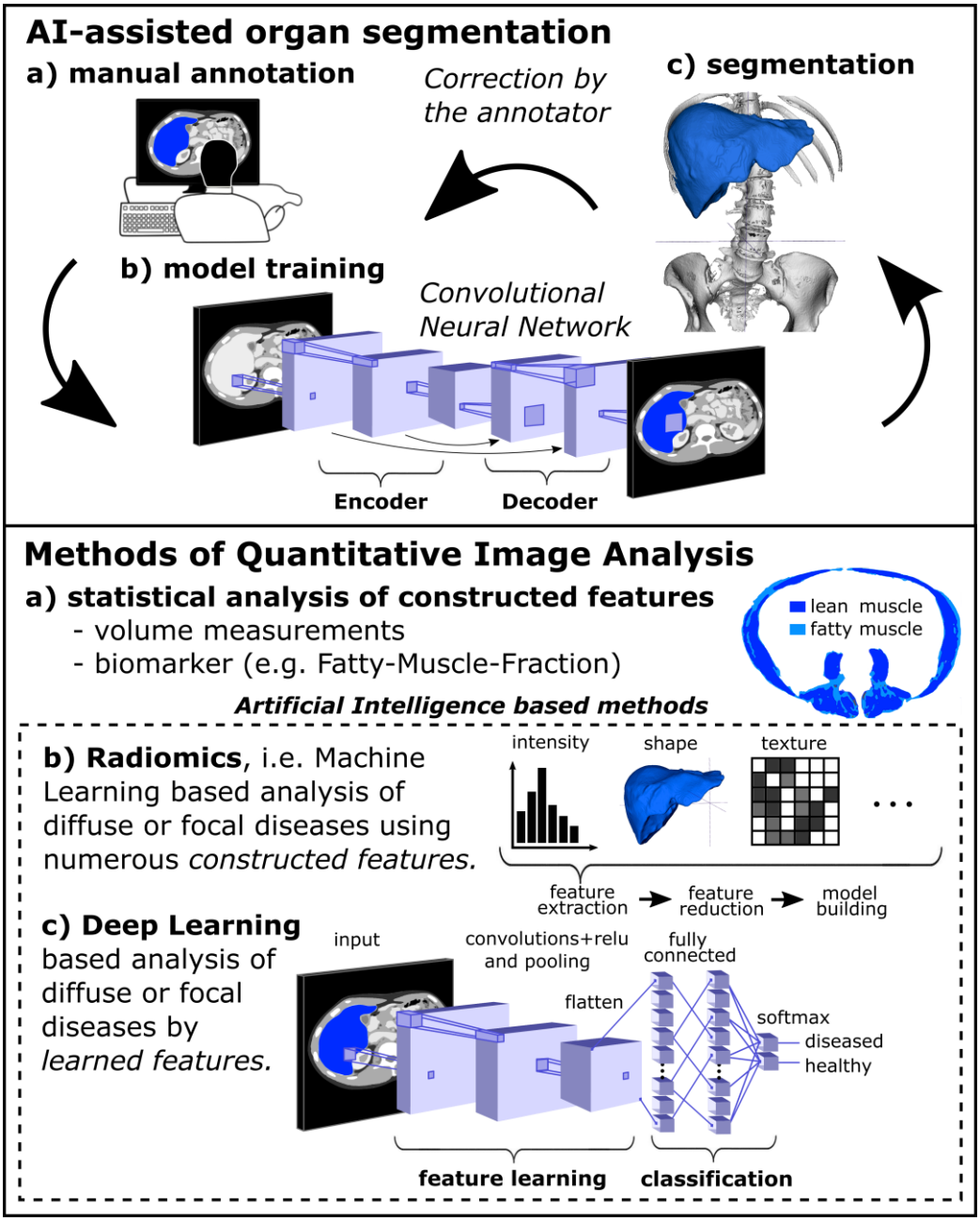


Figure 2.1: Schematic illustration of Quantitative Image Analysis pipeline development. Often, the development of automated segmentation of the tissues of interest is needed, which requires tedious manual annotations. Here, concepts such as AI-assisted segmentation can be beneficial, where an early and imperfect Convolutional Neural Network (CNN) that was trained with a small number of manually segmented images is applied to segment further images in advance, requiring less time-consuming optimization. Subsequently, user-defined 'constructed' features can be analysed for their diagnostic or prognostic value using a) conventional statistical analyses such as Kaplan-Meier analyses or b) in Machine Learning based Radiomics analyses. c) When employing Deep Learning methods and especially CNNs, the image features are usually not user-defined, but learned and recognized by the method itself.

References

Faron A, Opheys NS, Nowak S, Sprinkart AM, Isaak A, Theis M, Mesropyan N, Endler C, Sirokay J, Pieper CC, Kuetting D, Attenberger UI, Landsberg L, Luetkens JA. Deep Learning-Based Body Composition Analysis Predicts Outcome in Melanoma Patients Treated with Immune Checkpoint Inhibitors. Diagnostics 2021; 11(12): 2314

Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2015; 278(2): 563–577

Hagiwara A, Fujita S, Ohno Y, Aoki S. Variability and standardization of quantitative imaging: monoparametric to multiparametric quantification, radiomics, and artificial intelligence. Invest Radiol 2020; 55(9): 601-616

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018; 18(8): 500-510

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal 2017; 42: 60–88

Luetkens JA, Faron A, Geissler HL, Al-Kassou B, Shamekhi J, Stundl A, Sprinkart AM, Meyer C, Fimmers R, Treede H, Grube E, Nickenig G, Sinning JM, Thomas D. Opportunistic computed tomography imaging for the assessment of fatty muscle fraction predicts outcome in patients undergoing transcatheter aortic valve replacement. Circulation 2020; 141:234–236

Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. Can J Cardiol 2022; 38(2): 204-213

Prado CMM, Lieffers JR, McCargar LJ, Reiman T, Sawyer MB, Martin L, Baracos VE. Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. Lancet Oncol 2008; 9: 629–635

3. Publications

Publication 1

Nowak S, Theis M, Wichtmann BD, Faron A, Froelich MF, Tollens F, Geißler HL, Block W, Luetkens JA, Attenberger UI, Sprinkart AM. **End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT.** Eur Radiol 2022; 32: 3142–3151, DOI: 10.1007/s00330-021-08313-x

Publication 2

Nowak S, Mesropyan N, Faron A, Block W, Reuter M, Attenberger UI, Luetkens JA, Sprinkart AM. **Detection of liver cirrhosis in standard T2-weighted MRI using deep transfer learning.** Eur Radiol 2021; 31: 8807–8815, DOI: 10.1007/s00330-021-07858-1

Publication 3

Nowak S, Henkel A, Theis M, Luetkens JA, Geiger S, Sprinkart AM, Pieper CC, Attenberger UI. **Deep learning for standardized, MRI-based quantification of subcutaneous and subfascial tissue volume for patients with lipedema and lymphedema.** Eur Radiol 2023; 33: 884–892, DOI: 10.1007/s00330-022-09047-0

European Radiology (2022) 32:3142–3151
https://doi.org/10.1007/s00330-021-08313-x

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT

Sebastian Nowak[1] · Maike Theis[1] · Barbara D. Wichtmann[1] · Anton Faron[1] · Matthias F. Froelich[2] · Fabian Tollens[2] · Helena L. Geißler[1] · Wolfgang Block[1,3,4] · Julian A. Luetkens[1] · Ulrike I. Attenberger[1] · Alois M. Sprinkart[1]

Received: 15 April 2021 / Revised: 6 August 2021 / Accepted: 31 August 2021 / Published online: 30 September 2021

Abstract

Objectives To develop a pipeline for automated body composition analysis and skeletal muscle assessment with integrated quality control for large-scale application in opportunistic imaging.

Methods First, a convolutional neural network for extraction of a single slice at the L3/L4 lumbar level was developed on CT scans of 240 patients applying the nnU-Net framework. Second, a 2D competitive dense fully convolutional U-Net for segmentation of visceral and subcutaneous adipose tissue (VAT, SAT), skeletal muscle (SM), and subsequent determination of fatty muscle fraction (FMF) was developed on single CT slices of 1143 patients. For both steps, automated quality control was integrated by a logistic regression model classifying the presence of L3/L4 and a linear regression model predicting the segmentation quality in terms of Dice score. To evaluate the performance of the entire pipeline end-to-end, body composition metrics, and FMF were compared to manual analyses including 364 patients from two centers.

Results Excellent results were observed for slice extraction (z-deviation = 2.46 ± 6.20 mm) and segmentation (Dice score for SM = 0.95 ± 0.04, VAT = 0.98 ± 0.02, SAT = 0.97 ± 0.04) on the dual-center test set excluding cases with artifacts due to metallic implants. No data were excluded for end-to-end performance analyses. With a restrictive setting of the integrated segmentation quality control, 39 of 364 patients were excluded containing 8 cases with metallic implants. This setting ensured a high agreement between manual and fully automated analyses with mean relative area deviations of ΔSM = $3.3 \pm 4.1\%$, ΔVAT = $3.0 \pm 4.7\%$, ΔSAT = $2.7 \pm 4.3\%$, and ΔFMF = $4.3 \pm 4.4\%$.

Conclusions This study presents an end-to-end automated deep learning pipeline for large-scale opportunistic assessment of body composition metrics and sarcopenia biomarkers in clinical routine.

Key Points

- *Body composition metrics and skeletal muscle quality can be opportunistically determined from routine abdominal CT scans.*
- *A pipeline consisting of two convolutional neural networks allows an end-to-end automated analysis.*
- *Machine-learning-based quality control ensures high agreement between manual and automatic analysis.*

Keywords Body composition · Tomography, X-ray computed · Deep learning · Quality control · Sarcopenia

Abbreviations

CDFNet	Competitive dense fully connected network
CNN	Convolutional neural network
FMF	Fatty muscle fraction
SAT	Subcutaneous adipose tissue
SM	Skeletal muscle
VAT	Visceral adipose tissue

Introduction

Body composition analyses aim to determine the quantity of connective tissue compartments. In addition to quantifying the amount of adipose and muscle tissue, recent work proposed methods to obtain additional information about a patient's general condition by also determining the quality of skeletal muscle tissue in terms of fatty degeneration. Several studies demonstrated that these metrics

Sebastian Nowak and Maike Theis contributed equally to this study.

✉ Alois M. Sprinkart
sprinkart@uni-bonn.de

Extended author information available on the last page of the article

determined from abdominal imaging provide prognostic implications in oncologic or cardiovascular diseases [1–8].

The amount of visceral and subcutaneous adipose tissue, as well as the amount and quality of muscle tissue, can be reliably determined from abdominal CT imaging. An opportunistic large-scale assessment in clinical routine has the potential to further enhance the understanding of the clinical value of body composition analyses in various diseases, e.g., for therapy decision and/or outcome prediction. Also, the establishment of gender-, age-, and ethnicity-specific norm values is only feasible through the widespread application of these analyses.

However, the determination of fat and muscle volume by manually annotating the region of interest by a radiologist is rather time-consuming, which currently prevents clinical routine application. Several studies have shown that area measurements of connective tissue compartments on a single slice at a certain lumbar level are highly correlated with total volume in the abdomen [9–11]. This led to greatly reduced annotation times for manual body composition analysis when applying a 2D— instead of a 3D approach. In recent years, several methods have been proposed for automating the required tissue segmentation step. It was a logical consequence that with the dominant rise of deep learning for image segmentation the previously manually segmented images were used to develop methods for automated segmentation by supervised learning [12–14]. However, manual interaction was still required for extraction of the single slice on which the automatic segmentation is performed. Only very recent work also includes deep-learning-based automated slice extraction as the next step for truly automated body composition analyses [15–17].

Moreover, to the best of our knowledge, there is currently no work that presents integrated quality control for both slice extraction and tissue segmentation. This still leaves one factor that represents an additional human effort in opportunistic analysis, namely identifying cases where the algorithm fails. Automatic determination of the predictive uncertainties can help identify cases with low-quality analyses and can additionally be used to monitor the performance of an autonomous system during deployment, as suggested for machine learning operations to manage deep learning life cycles. This can also help to detect changes in the data and to raise a warning in case of domain shifts.

Hence, the aim of this study was to develop an automated body composition analysis for abdominal CT with integrated quality checks and to evaluate the end-to-end performance of the proposed pipeline on dual-center test data.

Material and methods

Overview

Figure 1 shows an overview of the developed pipeline. In the first part, a single slice at the L3/L4 lumbar level is extracted from a 3D CT scan. In the second part, the extracted 2D image is segmented into three compartment classes: visceral and subcutaneous adipose tissue (VAT, SAT) and skeletal muscle (SM). The fatty muscle fraction (FMF), a quantitative marker for fatty muscle degeneration, is determined in a subsequent post-processing step [1, 6]. For both deep-learning-based slice extraction and segmentation, classical machine learning methods were employed for integration of quality control steps that capture the predictive uncertainty during deployment.

Slice extraction and tissue segmentation were developed independently. To evaluate the end-to-end performance of the entire pipeline, automatically extracted body composition metrics and FMF were compared with manual analyses on an unselected dual-center test set. Figure 2 provides an overview of the data sets used for method development and evaluation.

Method development for slice extraction

Dataset

With institutional review board approval, written informed patient consent was waived because of the retrospective nature of all parts of the study. Retrospectively derived 3D CT scans of 240 patients (94 female, mean age 65 ± 14 years) referred for diagnostic CT including imaging of the upper abdomen acquired at eight different CT scanners were used for development of the slice extraction method. Of these patients, 43 received CT before undergoing transcatheter aortic valve implantation, 91 before transjugular intrahepatic portosystemic shunt intervention, and 106 patients received CT in the setting of immunotherapy for malignant melanoma.

The ground truth was generated by a board-certified radiologist (A.F.) by manually defining the center of the L3/L4 vertebral disk with an in-house tool (Matlab, Mathworks). Data were randomly split into a training set ($n = 192$, 80%) and a hold-out test ($n = 48$, 20%) set. The method was additionally tested on dual-center test data (described below).

Model

The extraction of a single slice at L3/L4 lumbar level was formulated as a segmentation task. A 3D U-Net architecture was trained using the nnU-Net framework, which has achieved high-performance values for various medical segmentation tasks and has the advantage of automatically adapting to different input sizes [18]. This is a relevant feature for the slice extraction task since the input are CT scans with a wide variety of scan lengths. The label map for

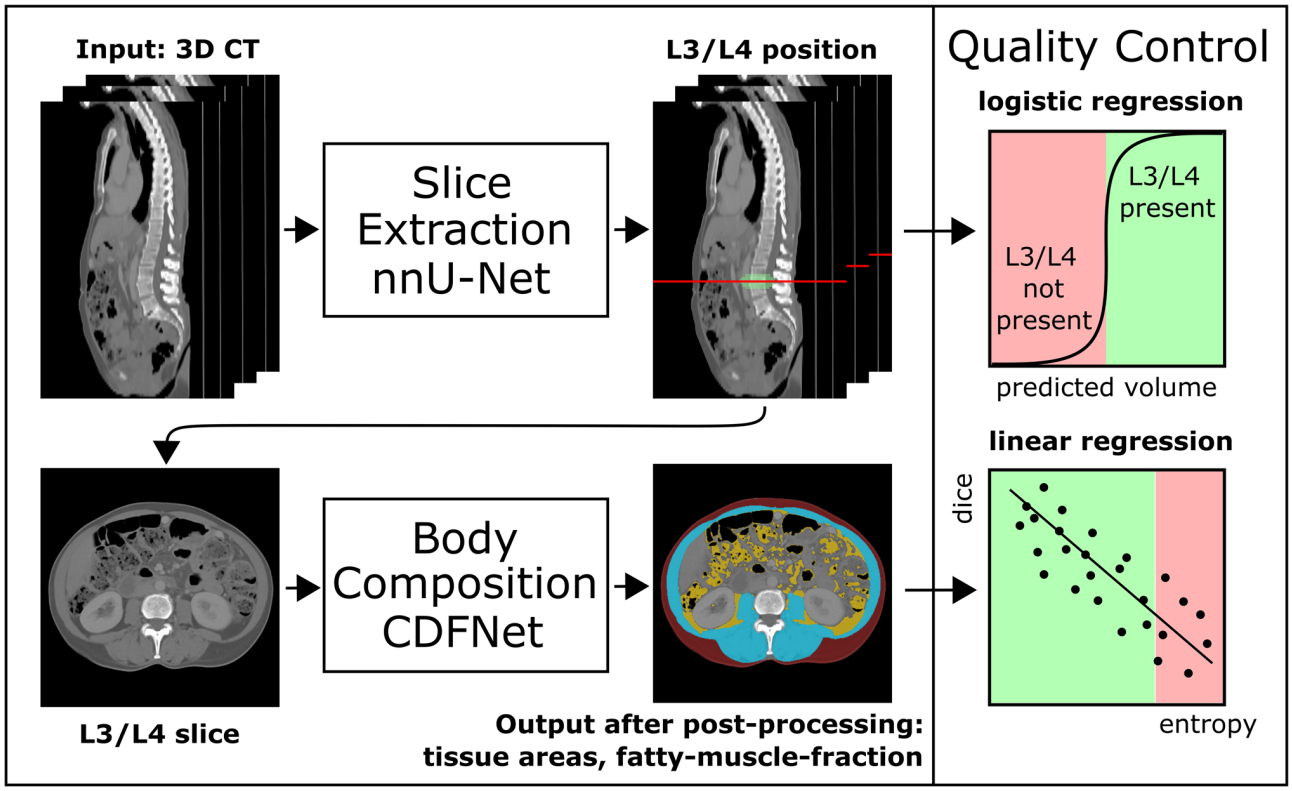


Fig. 1 Schematic representation of the presented pipeline for autonomous body composition analysis. Input of the pipeline is a 3D CT scan. In the first part, a 3D convolutional neural network (CNN) was employed for slice extraction using nnU-Net. In the second part, a competitive dense fully connected CNN (CDFNet) is applied for segmentation of the body compartments. Classical machine learning methods were employed for integration of quality control steps. For the slice extraction part, a logistic regression model was developed that classifies the presence of L3/L4 lumbar level in the 3D CT scan. For segmentation of the different tissues, a linear regression model was established that predicts segmentation quality in terms of the Dice score

training of the network was generated by applying a Gaussian distribution to the coordinates of the L3/L4 vertebral disk and binarizing the resulting probability map by a threshold [19]. Further details on image pre-processing, augmentation, and experimental design can be found in Supplement S1. For training, fivefold cross-validation was used and testing was performed with an ensemble of the cross-validated models.

Quality control

After training of the slice extraction method, a logistic regression model was built to automatically identify 3D CT scans that do not include the L3/L4 lumbar level. To obtain a balanced distribution of images with and without the L3/L4 lumbar level, for each 3D CT scan of the training, hold-out and dual-center test set, a cropped version was created. The logistic regression model was trained based on the predicted volume of all validation cases of the cross-validated slice extraction nnU-Net and applied to all test sets. Additional information about cropping and feature selection can be found in Supplement S2.

Method development for tissue segmentation

Dataset

For the development of the tissue segmentation method (VAT, SAT, SM), retrospectively derived single slice images at the L3/L4 lumbar level from 1143 patients (559 female, mean age 77 ± 11 years) were used. 937 patients underwent pre-interventional CT for transcatheter aortic valve implantation and 206 patients underwent diagnostic CT for liver cirrhosis with portosystemic shunting. The dataset intentionally included a high number of patients with anasarca (19.2%), ascites (9.4%), or both anasarca and ascites (6.5%). The ground truth of the segmentation was defined by manual drawing and was also used to train a different CNN in a previous work, where additional details on the dataset are reported [13].

The data for method development were randomly split into a training set ($n = 972$, 85%) and hold-out test ($n = 171$, 15%) set. The method was additionally tested on dual-center test data (described below).

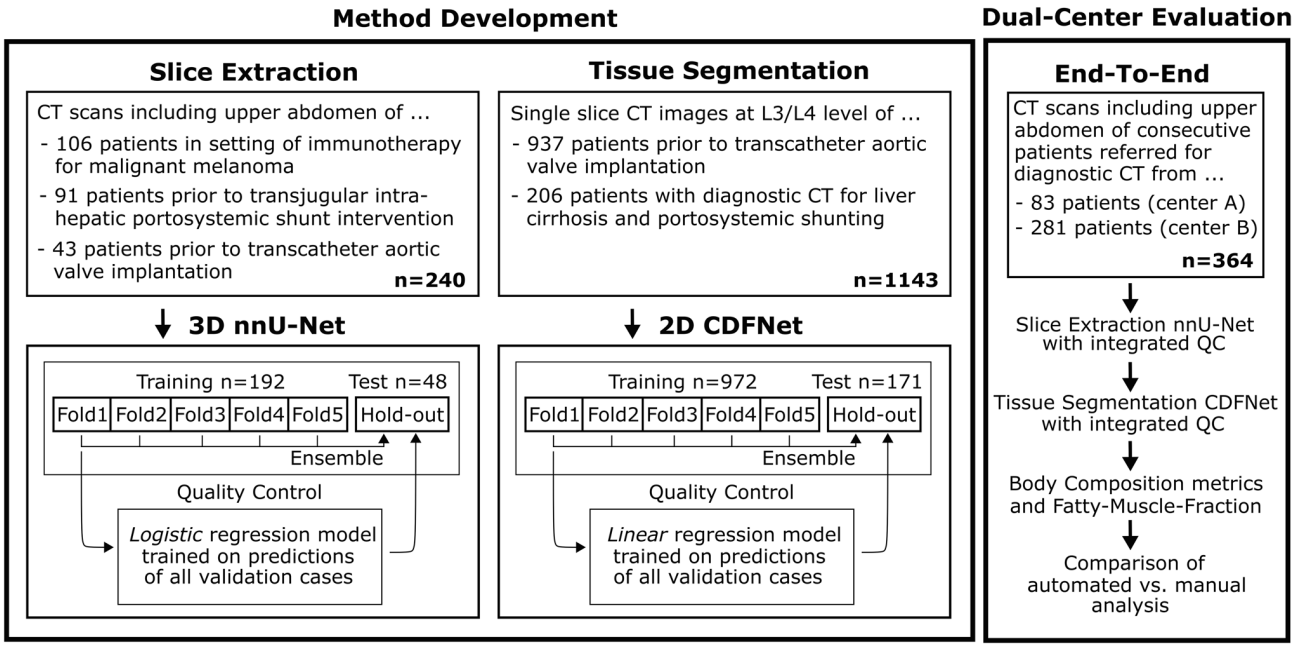


Fig. 2 Overview of the data sets used for method development and evaluation. The nnU-Net employed for extraction of a single slice at L3/L4 level from a 3D CT scan and the CDFNet for tissue segmentation of the 2D CT slices were developed on two different datasets. Both methods were fivefold cross-validated and an ensemble of the cross-validated models was tested on the hold-out data. The regression models for integrated quality control (QC) were developed on the validation data of the cross-validated models and were also tested on the hold-out data. Finally, the entire pipeline of slice extraction, tissue segmentation, and quality control was evaluated end-to-end on the dual-center test data and compared against manual analyses

Model

A 2D competitive dense fully convolutional network (CDFNet), which has shown promising results for body composition analysis in magnetic resonance imaging, was used for tissue segmentation [20]. This architecture is proposed as an extension of the Dense-UNet architecture by max-out activation units. In a CDFNet, feature maps are generated by element-wise selection of the maximum values of previous feature maps, which has been shown to have a positive effect on performance and generalizability compared to unselective concatenation [20–22]. Further details on image pre-processing, augmentation, experimental design and computation of the fatty muscle fraction are provided in Supplement S3.

For training, fivefold cross-validation was used and testing was performed with an ensemble of the cross-validated models.

Quality control

To assess the predictive uncertainty of the segmentation during employment, a linear regression model was developed that predicts the segmentation Dice score for the muscle class based on the average entropies of the probability maps. This metric is proposed by a recent work as a feature to estimate quality of medical image segmentation and to detect out-of-distribution samples and ambiguous cases [23]. Although this method could be applied to all tissue classes, we focused on the muscle class because we consider it the most important class for the assessment of sarcopenia.

The linear regression model was trained with the predictions of all validation cases of the cross-validated tissue segmentation CDFNet and tested on all test sets.

Dual-center test data and end-to-end evaluation

The entire pipeline was finally evaluated end-to-end, i.e., from 3D CT scan to extracted body composition metrics. The automatically determined tissue areas and the fatty muscle fraction were compared with the manually determined values. For this purpose, 3D CT scans of consecutive patients referred for diagnostic CT including imaging of the upper abdomen were retrospectively retrieved from two centers.

- Center A: 83 (41 females, mean age 60 ± 15 years) patients were used as internal test data from the Department of Diagnostic and Interventional Radiology, Uni-

versity Hospital Bonn. Data were acquired at four different CT scanners.

- Center B: 281 (111 females, mean age 63 ± 16 years) patients were used as external test data from the Department of Radiology and Nuclear Medicine, University Medical Centre Mannheim. Data were acquired at three different CT scanners.

In this data set, 10 patients had metallic implants. However, in the end-to-end evaluation, these cases were intentionally not excluded. For demonstration of the tissue segmentation quality control, a restrictive setting was applied excluding 10% of the cases with lowest predicted Dice score of the muscle class. End-to-end performance is reported for both included and excluded cases.

The ground truth for slice extraction and tissue segmentation was labeled by a radiology resident (B.W.) and a board-certified radiologist (A.F.). All labels of the radiology resident were validated by the board-certified radiologist.

Additional information on dual-center test data can be found in Supplement S5.

Results

A summary of the results can be found in Fig. 3.

Slice extraction

The mean deviations between the predictions of the ensemble of cross-validated slice extraction models and the manually defined ground truth were $\Delta z = 2.27 \pm 7.08$ mm for the hold-out test data and $\Delta z = 2.46 \pm 6.20$ mm for the dual-center test data. Considering an acceptable deviation of up to 10 mm, 96% of the extracted slices of the hold-out test set and 96% of the dual-center test data were extracted at the correct level. The mean deviations are listed separately for all test sets in Table 1.

Tissue segmentation

The ensemble of fivefold cross-validated CDFNet models achieved excellent Dice scores, both on the hold-out test data (SM: 0.96 ± 0.02, VAT: 0.98 ± 0.02, SAT: 0.98 ± 0.01) and on the dual-center test data (SM: 0.95 ± 0.04, VAT: 0.98 ± 0.02, SAT: 0.97 ± 0.04). Table 2 lists the Dice scores separately for each test set.

Quality control

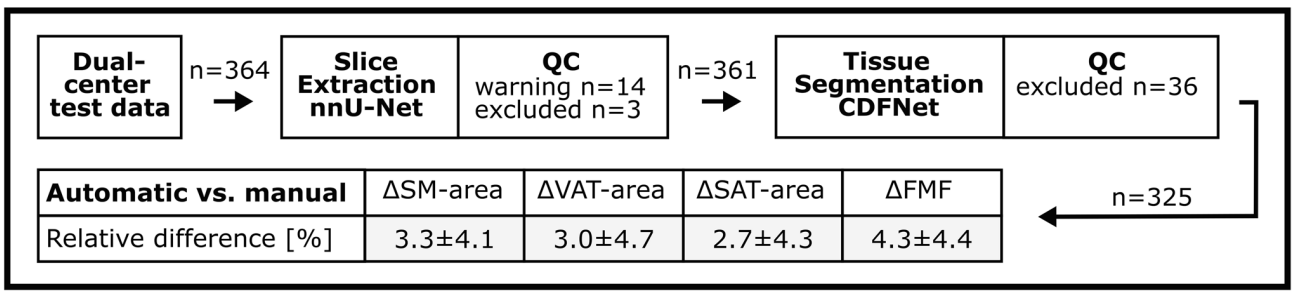
Figure 4a shows the logistic regression model developed for identifying 3D CT scans that do not contain the L3/L4 level. High accuracy was observed for predicting the presence of

Seperate Evaluation of Slice Extraction and Tissue Segmentation

Slice Extraction			QC
Results	Mean Δz [mm]	Accuracy Δz<=10mm	Accuracy L3/L4 present
Hold-out test data	2.3±7.1	0.96	1.00
Dual-center test data	2.5±6.2	0.96	0.98

Tissue Segmentation				QC
Results	Dice SM	Dice VAT	Dice SAT	ΔDice SM predicted
Hold-out test data	0.96	0.98	0.98	0.016
Dual-center test data	0.95	0.98	0.97	0.016

End-To-End Evaluation



Automatic vs. manual	ΔSM-area	ΔVAT-area	ΔSAT-area	ΔFMF
Relative difference [%]	3.3±4.1	3.0±4.7	2.7±4.3	4.3±4.4

Fig. 3 Summary of results: separate analyses of slice extraction, tissue segmentation, and respective quality control (QC), as well as agreement between end-to-end automated and manual area measurements of skeletal muscle (SM), visceral adipose tissue (VAT), subcutaneous adipose tissue (SAT), and the fatty muscle fraction (FMF)

Table 1 Mean z-deviation (Δz) and slice extraction accuracy for different tolerance margins obtained with the cross-validated nnU-Net ensemble for the hold-out test set and for the additional test data from center A and center B

Slice extraction	Mean, Δz [mm]	Accuracy, $\Delta z = 0$ mm	Accuracy, $\Delta z <= 5$ mm	Accuracy, $\Delta z <= 10$ mm
Hold-out	2.27 ± 7.08	0.79	0.96	0.96
Center A	3.35 ± 4.10	0.51	0.88	0.99
Center B	2.19 ± 6.70	0.85	0.96	0.96

Table 2 Dice scores for segmentation of skeletal muscle (SM), visceral adipose tissue (VAT), and subcutaneous adipose tissue (SAT) obtained with the cross-validated CDFNet ensemble for the hold-out test set and for the additional test data from center A and center B

Tissue segmentation	Dice score, SM	Dice score, VAT	Dice score, SAT
Hold-out	0.958 ± 0.023	0.981 ± 0.015	0.982 ± 0.012
Center A	0.959 ± 0.021	0.981 ± 0.012	0.979 ± 0.038
Center B	0.944 ± 0.039	0.974 ± 0.027	0.969 ± 0.037

the L3/L4 level in the original and cropped versions of the hold-out test data (100%) and also on the dual-center test data (center A: 99%, center B: 98%). Sensitivity and specificity were 97% and 99% for the dual-center test data.

The linear regression model developed for integrated quality control of the tissue segmentation is shown in Fig. 4b. Mean differences between the observed and the predicted Dice score for the hold-out test data were 0.016 ± 0.016 (SM), 0.005 ± 0.005 (VAT), and 0.008 ± 0.010 (SAT) and for the dual-center 0.016 ± 0.016 (SM), 0.007 ± 0.012 (VAT), and 0.010 ± 0.015 (SAT).

End-to-end evaluation

Figure 5 shows examples of the end-to-end analyses. Application of the logistic regression model to the dual-center test data, all of which contained the L3/L4 lumbar level, resulted in 14 of 364 3D CT scans with a warning that the scan may not contain the L3/L4 level. In three of these cases, the patients had implants at the L3/L4 level. For the remaining 11 cases, the difference between predicted L3/L4 level and ground truth was $\Delta z = 6.38 \pm 10.77$ mm. Except for the three patients with implants, none of the patients were excluded from further analyses. Subsequently, the linear regression model for integrated quality control of the tissue segmentation was applied. With a restrictive setting, 36 of 361 cases were flagged as possibly having limited segmentation quality with predicted Dice scores of the muscle class ranging from 0.861 to 0.924. In 5 of these 36 cases, the patients had implants at the L3/L4 level, and 4 patients had a pronounced hernia. In the remaining cases, there were various reasons for limited segmentation quality, such as parts of the arms included in the tissue segmentation or parts of the kidney classified as muscle. In total, 8 of 10 cases with metallic implants on the L3/L4 level were excluded by the two quality control steps. For the two cases not excluded by quality

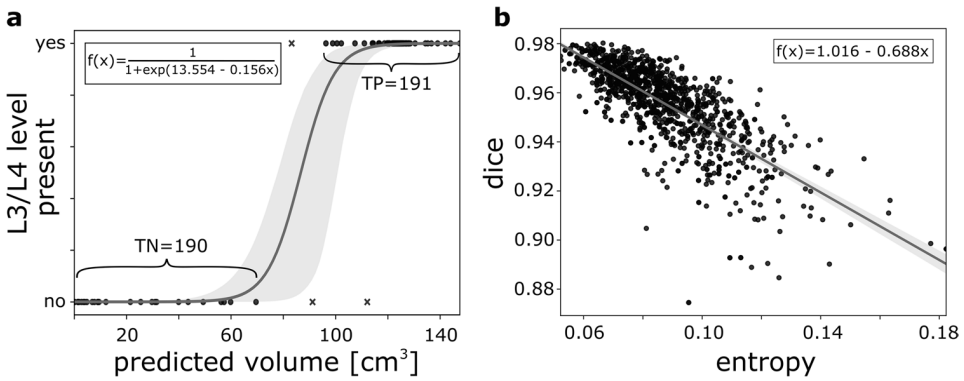


Fig. 4 Models trained for quality control: **a** Based on the predicted volume of the nnU-Net employed for slice extraction, a logistic regression model was trained to predict the presence of the slice at L3/L4 lumbar level in the 3D CT scan. **b** For prediction of the tissue segmentation quality in terms of the Dice score, a linear regression model was trained based on the entropy of the probability map of the CDFNet for the muscle class. Both regression models were built on features derived from cross-validation data of slice extraction and tissue segmentation, respectively. Gray areas represent the 95% confidence intervals

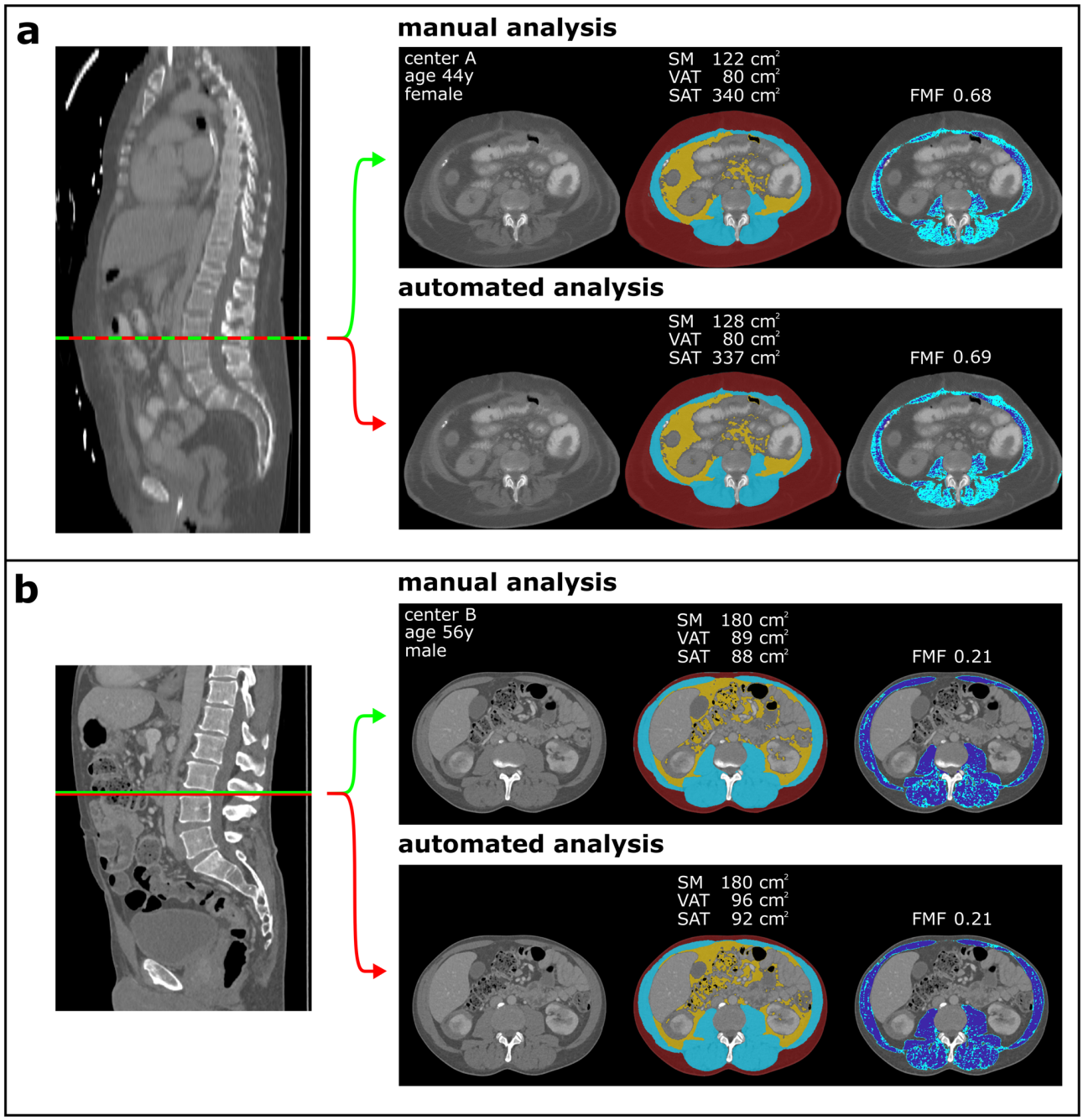


Fig. 5 Compartmental areas of visceral adipose tissue, subcutaneous adipose tissue (VAT, SAT), skeletal muscle (SM), and fatty muscle fraction (FMF) derived for patients from center A (**a**) and center B (**b**). Manual analysis is marked in green, while results from the proposed pipeline are marked with a red line

control, only minor hardening artifacts were observed, as shown in Supplement 4S.

Results of the entire end-to-end evaluation are summarized in Table 3. A high agreement was observed for the 325 cases of the dual-center data that passed the quality control. Body composition metrics and FMF derived from automated and manual analysis showed absolute differences in area of $\Delta SM = 5.0 \pm 6.0$ cm^2, $\Delta VAT = 3.7 \pm 5.8$ cm^2, and $\Delta SAT = 5.7 \pm 10.4$ cm^2, corresponding to low relative differences of $\Delta SM = 3.3 \pm 4.1\%$, $\Delta VAT = 3.0 \pm 4.7\%$, and $\Delta SAT = 2.7 \pm 4.3\%$. Also for FMF, low absolute deviations of $\Delta FMF = 0.014 \pm 0.012$ and relative deviations of $\Delta FMF = 4.3 \pm 4.4\%$ were observed.

Table 3 Evaluation of the end-to-end performance of the body composition analyses

Center	Quality control	Fatty muscle fraction	Muscle area (cm^2)	Visceral fat area (cm^2)	Subcutaneous fat area (cm^2)
A	Passed, $n=82$	0.009 ± 0.008 (3.1% ± 3.5%)	3.7 ± 4.1 (2.7% ± 4.4%)	3.6 ± 4.3 (2.7% ± 3.6%)	5.4 ± 5.3 (2.7% ± 3.0%)
B	Passed, $n=243$	0.016 ± 0.013 (4.8% ± 4.6%)	5.4 ± 6.4 (3.5% ± 4.0%)	3.8 ± 6.2 (3.1% ± 5.0%)	5.8 ± 11.7 (2.8% ± 4.6%)
A	Excluded, $n=1$	**0.046 (9.3%)**	**16.0 (16.6%)**	2.0 (2.3%)	14.9 (10.8%)
B	Excluded, $n=35$	**0.033 ± 0.036 (6.1% ± 6.6%)**	**18.6 ± 21.6 (14.1% ± 15.6%)**	7.2 ± 10.4 (7.0% ± 8.6%)	18.4 ± 29.5 (7.8% ± 9.5%)

Absolute and relative differences (in parentheses) between the values obtained with the proposed pipeline and the manually determined values are listed separately for center A and center B and for all 3D CT scans that were included and excluded by restrictive setting of the tissue segmentation quality control. The excluded cases show markedly lower agreement of muscle area, while FMF agreement is still reasonably good (marked in bold)

Discussion

This paper presents a method that allows the application of body composition analysis without human interaction, thus permitting opportunistic determination of body compartment areas and FMF as a marker for sarcopenia in routine clinical practice. For both CNNs applied in the pipeline, the trained networks are available on reasonable request (https://qilab.de).

In recent years, a variety of deep learning methods have been presented that address the topic of automated body composition analysis. Most of these studies focus on the segmentation of the tissue compartments in a single slice at a certain lumbar level, as it has been demonstrated that 2D and 3D measurements for quantification of VAT, SAT, and SM show a high correlation [9–14]. Although very recent works have also addressed automation of slice extraction, routine clinical application additionally requires the integration of quality control methods for both slice extraction and tissue segmentation [15, 16]. For this purpose, two classic machine learning models have been developed in this study. The developed pipeline therefore provides full automation of body composition analysis in abdominal CT, including deep-learning-based slice extraction and tissue segmentation and integrated application of quality control models.

Compared to previous research in the field of automated body composition analyses, we observed similar or superior performance values for slice extraction task and tissue segmentation in our study [12–17]. In previous work, the slice extraction task was formulated either as a regression problem, a classification task, or, similar to our approach, a segmentation problem [15–17]. While the methods proposed so far for slice extraction are based on 2D images or require the generation of a maximum intensity projection in a preprocessing step, the use of the nnU-Net framework allows the direct input of 3D CT datasets of different sizes. For tissue segmentation, different variants of a 2D U-Net architecture have been used [12, 15–17]. The CDFNet architecture applied in the current study is an extension of a DenseUNet architecture with max-out activation units, which has recently also been successfully used for body composition analyses in magnetic resonance imaging [20]. A detailed comparison to previous work can be found in Supplement S6.

For the development of the tissue segmentation CNN, patient collectives were included that also represent tissue alterations, as ascites and anasarca, which are challenging for body composition analysis [14]. In addition, segmentation results from other studies show the disadvantages of using only threshold-based pre-processing steps to define segmentation ground truth, resulting in misclassification of intermuscular fat to one of the abdominal adipose tissue classes (VAT, SAT) [15]. To overcome this limitation, intermuscular fat was manually assigned to the muscle class in this study, allowing additional analyses of muscle [13].

Several aspects of body composition, such as skeletal muscle fat infiltration as an indicator of skeletal muscle quality were shown to provide prognostic information in patients with cardiovascular and oncologic diseases [1–3]. Thereby, FMF was recently proposed as an easy-accessible body composition metric which may be considered particularly promising as it additionally integrates information on skeletal muscle quality [1, 5]. Previous studies have demonstrated its prognostic value both as an indicator of frailty in patients with planned endovascular aortic valve replacement as well as an powerful predictor of outcome in critically ill patients receiving extracorporeal membrane oxygenation therapy [1, 6].

A recent work on 3D tissue segmentation points out that for a truly automated application of body compartment analysis, the development of quality assurance procedures is warranted to identify patients with metal artifacts [24]. The dual-center end-to-end analysis presented in the current work demonstrates that the proposed quality control ensures a high agreement between manual and automated analyses by identifying cases that are unsuitable for body composition analyses not only due to hardening artifacts but also due to other reasons limiting the segmentation quality. Interestingly, end-to-end performance analysis of cases flagged by quality control as having limited segmentation quality shows that FMF is quite robust to segmentation errors.

As a limitation of this study, only the areas of VAT, SAT, and SM are determined in a single slice instead of determining the respective tissue volumes in the entire abdomen. However, we are not aware of studies demonstrating that a 3D approach has significant advantages over the established 2D measurement for assessment of sarcopenia. Also, reference values for body compartments have so far only been determined in large studies for 2D measurements [15].

Conclusion

This study presents an end-to-end automated deep-learning pipeline for large-scale opportunistic assessment of body composition metrics and sarcopenia biomarker in clinical routine.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-08313-x.

Funding Open Access funding enabled and organized by Projekt DEAL. The study was supported by a grant from the BONFOR research program of the University of Bonn (application number 2020-2A-04). The funders had no influence on conceptualization and design of the study, data analysis, and data collection, preparation of the manuscript as well as the decision to publish.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is PD Dr. Alois Martin Sprinkart.

Conflict of interest The authors declare no competing interests.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was waived by the institutional review board (the University of Bonn and the University of Heidelberg).

Ethical approval This retrospective study was approved by the institutional review board with waiver of written informed consent.

Methodology

retrospective
diagnostic study
performed at two institutions

References

1. Luetkens JA, Faron A, Geissler HL et al (2020) Opportunistic computed tomography imaging for the assessment of fatty muscle fraction predicts outcome in patients undergoing transcatheter aortic valve replacement. Circulation 141:234–236
2. Faron A, Pieper CC, Schmeel FC et al (2019) Fat-free muscle area measured by magnetic resonance imaging predicts overall survival of patients undergoing radioembolization of colorectal cancer liver metastases. Eur Radiol 29:4709–4717
3. Faron A, Sprinkart AM, Pieper CC, et al (2020) Yttrium-90 radioembolization for hepatocellular carcinoma: outcome prediction with MRI derived fat-free muscle area. Eur J Radiol 125:108889.
4. Faron A, Sprinkart AM, Kuetting DLR et al (2020) Body composition analysis using CT and MRI: intra-individual intermodal comparison of muscle mass and myosteatosis. Sci Rep 10:11765
5. Cruz-Jentoft AJ, Bahat G, Bauer J et al (2019) Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 48:16–31
6. Faron A, Kreyer S, Sprinkart AM et al (2020) CT fatty muscle fraction as a new parameter for muscle quality assessment predicts outcome in venovenous extracorporeal membrane oxygenation. Sci Rep 10:22391
7. Lenchik L, Boutin RD (2018) Sarcopenia: beyond muscle atrophy and into the new frontiers of opportunistic imaging, precision medicine, and machine learning. Semin Musculoskelet Radiol 22:307–322
8. Prado CMM, Lieffers JR, McCargar LJ et al (2008) Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. Lancet Oncol 9:629–635
9. Shen W, Punyanitya M, Wang Z et al (2004) Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. J Appl Physiol 97:2333–2338
10. Faron A, Luetkens JA, Schmeel FC et al (2019) Quantification of fat and skeletal muscle tissue at abdominal computed tomography: associations between single-slice measurements and total compartment volumes. Abdom Radiol 44:1907–1916
11. Irlbeck T, Massaro JM, Bamberg F, O'Donnell CJ, Hoffmann U, Fox CS (2010) Association between single-slice measurements of visceral and abdominal subcutaneous adipose tissue with volumetric measurements: the Framingham Heart Study. Int J Obes (Lond) 34:781–787
12. Weston AD, Korfiatis P, Kline TL et al (2018) Automated abdominal segmentation of CT scans for body composition analysis using deep learning. Radiology 290:669–679
13. Nowak S, Faron A, Luetkens JA et al (2020) Fully automated segmentation of connective tissue compartments for CT-based body composition analysis: a deep learning approach. Invest Radiol 55:357–366
14. Park HJ, Shin Y, Park J et al (2020) Development and validation of a deep learning system for segmentation of abdominal muscle and fat on computed tomography. Korean J Radiol 21:88–100
15. Magudia K, Bridge CP, Bay CP et al (2020) Population-scale CT-based body composition analysis of a large outpatient population using deep learning to derive age-, sex-, and race-specific reference curves. Radiology 298:319–329

16. Dabiri S, Popuri K, Ma C, et al (2020) Deep learning method for localization and segmentation of abdominal CT. Comput Med Imaging Graph 85:101776.
17. Castiglione J, Somasundaram E, Gilligan LA, Trout AT, Brady S (2021) Automated segmentation of abdominal skeletal muscle on pediatric ct scans using deep learning. Radiol Artif Intell 3:e200130.
18. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18:203–211
19. Yang D, Xiong T, Xu D et al (2017) Deep Image-to-Image Recurrent Network with Shape Basis Learning for Automatic Vertebra Labeling in Large-Scale 3D CT Volumes. Proceedings of MICCAI 2017:498–506
20. Estrada S, Lu R, Conjeti S et al (2020) FatSegNet: A fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. Magn Reson Med 83:1471–1483
21. Estrada S, Conjeti S, Ahmad M, Navab N, Reuter M (2018) Competition vs. concatenation in skip connections of fully convolutional networks. Proceedings of international workshop on machine Learning in Medical Imaging, pp 214–222.
22. Goodfellow IJ, Warde-Farley D, Mirza M, Courville A, Bengio Y (2013) Maxout networks. Proceedings of International Conference on Machine Learning, pp 1319–1327.
23. Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T (2020) Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans Med Imag 39:3868–3878
24. Koitka S, Kroll L, Malamutmann E, Oezcelik A, Nensa F (2021) Fully automated body composition analysis in routine CT imaging using 3D semantic segmentation convolutional neural networks. Eur Radiol 31:1795–1804

Authors and Affiliations

Sebastian Nowak[1] · Maike Theis[1] · Barbara D. Wichtmann[1] · Anton Faron[1] · Matthias F. Froelich[2] · Fabian Tollens[2] · Helena L. Geißler[1] · Wolfgang Block[1,3,4] · Julian A. Luetkens[1] · Ulrike I. Attenberger[1] · Alois M. Sprinkart[1]

[1] Department of Diagnostic and Interventional Radiology, Quantitative Imaging Lab Bonn (QILaB), University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

[2] Department of Radiology and Nuclear Medicine, University Medical Centre Mannheim, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany

[3] Department of Radiotherapy and Radiation Oncology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

[4] Department of Neuroradiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

European Radiology (2021) 31:8807–8815
https://doi.org/10.1007/s00330-021-07858-1

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Detection of liver cirrhosis in standard T2-weighted MRI using deep transfer learning

Sebastian Nowak[1] · Narine Mesropyan[1] · Anton Faron[1] · Wolfgang Block[1] · Martin Reuter[2,3,4] · Ulrike I. Attenberger[1] · Julian A. Luetkens[1] · Alois M. Sprinkart[1]

Received: 11 November 2020 / Revised: 12 February 2021 / Accepted: 10 March 2021 / Published online: 11 May 2021

Abstract

Objectives To investigate the diagnostic performance of deep transfer learning (DTL) to detect liver cirrhosis from clinical MRI.

Methods The dataset for this retrospective analysis consisted of 713 (343 female) patients who underwent liver MRI between 2017 and 2019. In total, 553 of these subjects had a confirmed diagnosis of liver cirrhosis, while the remainder had no history of liver disease. T2-weighted MRI slices at the level of the caudate lobe were manually exported for DTL analysis. Data were randomly split into training, validation, and test sets (70%/15%/15%). A ResNet50 convolutional neural network (CNN) pre-trained on the ImageNet archive was used for cirrhosis detection with and without upstream liver segmentation. Classification performance for detection of liver cirrhosis was compared to two radiologists with different levels of experience (4th-year resident, board-certified radiologist). Segmentation was performed using a U-Net architecture built on a pre-trained ResNet34 encoder. Differences in classification accuracy were assessed by the χ^2-test.

Results Dice coefficients for automatic segmentation were above 0.98 for both validation and test data. The classification accuracy of liver cirrhosis on validation (vACC) and test (tACC) data for the DTL pipeline with upstream liver segmentation (vACC = 0.99, tACC = 0.96) was significantly higher compared to the resident (vACC = 0.88, $p < 0.01$; tACC = 0.91, $p = 0.01$) and to the board-certified radiologist (vACC = 0.96, $p < 0.01$; tACC = 0.90, $p < 0.01$).

Conclusion This proof-of-principle study demonstrates the potential of DTL for detecting cirrhosis based on standard T2-weighted MRI. The presented method for image-based diagnosis of liver cirrhosis demonstrated expert-level classification accuracy.

Key Points

- *A pipeline consisting of two convolutional neural networks (CNNs) pre-trained on an extensive natural image database (ImageNet archive) enables detection of liver cirrhosis on standard T2-weighted MRI.*
- *High classification accuracy can be achieved even without altering the pre-trained parameters of the convolutional neural networks.*
- *Other abdominal structures apart from the liver were relevant for detection when the network was trained on unsegmented images.*

Keywords Deep learning · Neural networks, computer · Magnetic resonance imaging · Liver cirrhosis

Sebastian Nowak and Narine Mesropyan contributed equally to this work.

✉ Alois M. Sprinkart
sprinkart@uni-bonn.de

[1] Department of Diagnostic and Interventional Radiology, Quantitative Imaging Lab Bonn (QILaB), University Hospital Bonn (Universitätsklinikum Bonn), Venusberg-Campus 1, 53127 Bonn, Germany

[2] Image Analysis, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[3] A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA

[4] Department of Radiology, Harvard Medical School, Boston, MA, USA

Abbreviations

ACC	Accuracy
AP	Average precision
AUC	Area under the curve
CNN	Convolutional neural network
DTL	Deep transfer learning

Introduction

Liver cirrhosis is the end stage of chronic liver disease and a major global health condition, especially due to its variety of severe complications caused by portal hypertension such as variceal bleeding, ascites, and hepatic encephalopathy [1]. Although liver biopsy is the gold standard for the detection of cirrhosis, imaging has a particularly important role in the evaluation of the disease [2]. Imaging is primarily used to characterize the morphologic manifestations of cirrhosis, evaluate the presence and the effects of portal hypertension, and screen for hepatocellular carcinoma. However, morphologic characteristics of cirrhosis are often detected incidentally in patients with unsuspected cirrhosis. It is therefore not unusual that radiologists presume an initial diagnosis of cirrhosis [3].

To assume a diagnosis of liver cirrhosis, different morphological criteria have been described for standard imaging modalities [2]. However, most of these findings are subjective, susceptible to inter-observer variability, and often lack high overall accuracy for the detection of cirrhosis [4]. Therefore, quantitative analyses, which could improve the objectivity and reading performance in the identification of liver cirrhosis, are of great interest [5].

A method that could objectively assess relevant features automatically within radiological images could support the radiologist in diagnosing liver cirrhosis, leading to greater accuracy and less variation in reading performance. Since 2012, when a deep learning technique has shown superior performance in the prominent ImageNet challenge for the first time, especially CNNs have become the gold standard for image classification and segmentation [6]. Deep learning methods have been continuously improved and successfully applied in various disciplines, including medical imaging [7–12].

However, a disadvantage of CNNs is the requirement of a large number of pre-classified images, which serve as training data. Instead of training a neural network from scratch with a small data set, it has proven advantageous to use a technique called transfer learning [13]. The basic idea is to use a CNN pre-trained e.g. on a large natural image dataset, which has already been trained to recognize complex patterns and then adapt it to a different task. This technique has recently been successfully applied to a variety of segmentations and classification problems of medical image data [14–16].

The aim of this study was to investigate the capabilities of deep transfer learning (DTL) to identify liver cirrhosis in standard T2-weighted MRI and to evaluate the diagnostic performance against radiologists with different levels of experience.

Materials and methods

This retrospective study was approved by the institutional review board with a waiver of written informed consent. Patients who underwent liver MRI at our institution for standard diagnostic purposes between 2017 and 2019 were included. Two groups of patients were identified and included in the final study cohort:

i. Patients with known liver cirrhosis of any stage: Inclusion criterion was the presence of histologically or clinically defined liver cirrhosis of any clinical disease severity. Exclusion criteria were the presence of focal liver lesions at the level of portal vein bifurcation or a past medical history of hepatic surgery (Fig. 1).
ii. Patients without known liver disease: From the same period, a randomly selected control group was recruited, which consisted of patients without known liver disease. Exclusion criteria for the control group were the same as those applied for the cirrhosis group.

Patient characteristics were retrieved from the clinical information management system of the referring institution. An overview of the MRI indications for the two groups is provided in Supplement S1.

As this study aimed to determine the diagnostic utility of DTL to detect liver cirrhosis based on morphological hallmarks of liver cirrhosis, T2-weighted imaging was used for analysis. In detail, images of a standard T2-weighted respiratory triggered multi-slice turbo spin echo sequence with non-Cartesian k-space filling with radial rectangular blades (Multi Vane XD) were used. For each patient, a single-slice image at the level of the caudate lobe was exported for DTL analysis (N.M. with 1 year of experience in the field of clinical abdominal imaging). All examinations were performed on clinical whole-body MRI systems (Philips, Ingenia 1.5 T and 3 T). Detailed imaging parameters are listed in Supplement S2.

Image data were randomly divided into training data (70%), validation data (15%), and test data (15%) using a custom Matlab script (MathWorks). Details of the preprocessing prior to training are listed in Supplement S3.

Images were analyzed using two different processing pipelines. In the first pipeline, an image segmentation network was applied prior to the classification task. In the second pipeline, the classification was performed directly on the unsegmented images.

For segmentation, a CNN following the principle architecture of a U-net model was implemented [17]. Its descending encoder part is identical to a CNN with residual connections

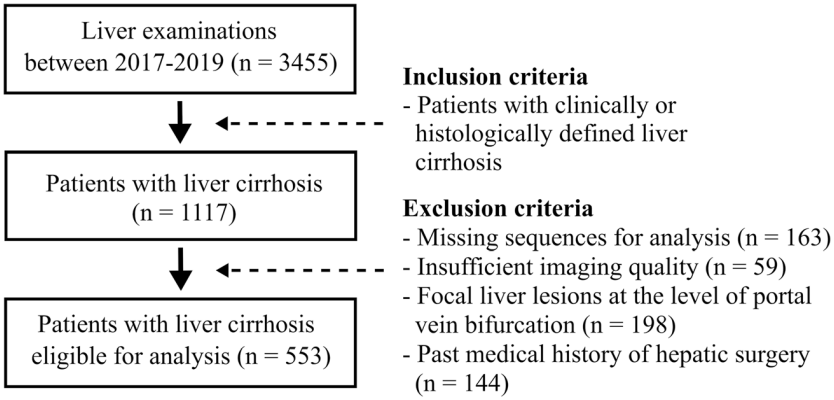


Fig. 1 Flowchart illustrating the inclusion and exclusion criteria for the group of patients with liver cirrhosis for this study

known as ResNet34 that was pre-trained on the ImageNet database [18]. The ground truth for the training of the segmentation CNN was generated by a radiology resident (N.M.) by manually delineating the liver using in-house tools developed in Matlab and verified by a board-certified radiologist (J.A.L.).

ResNet50 as a well-established CNN with 50 trainable layers and residual connections was used for the classification

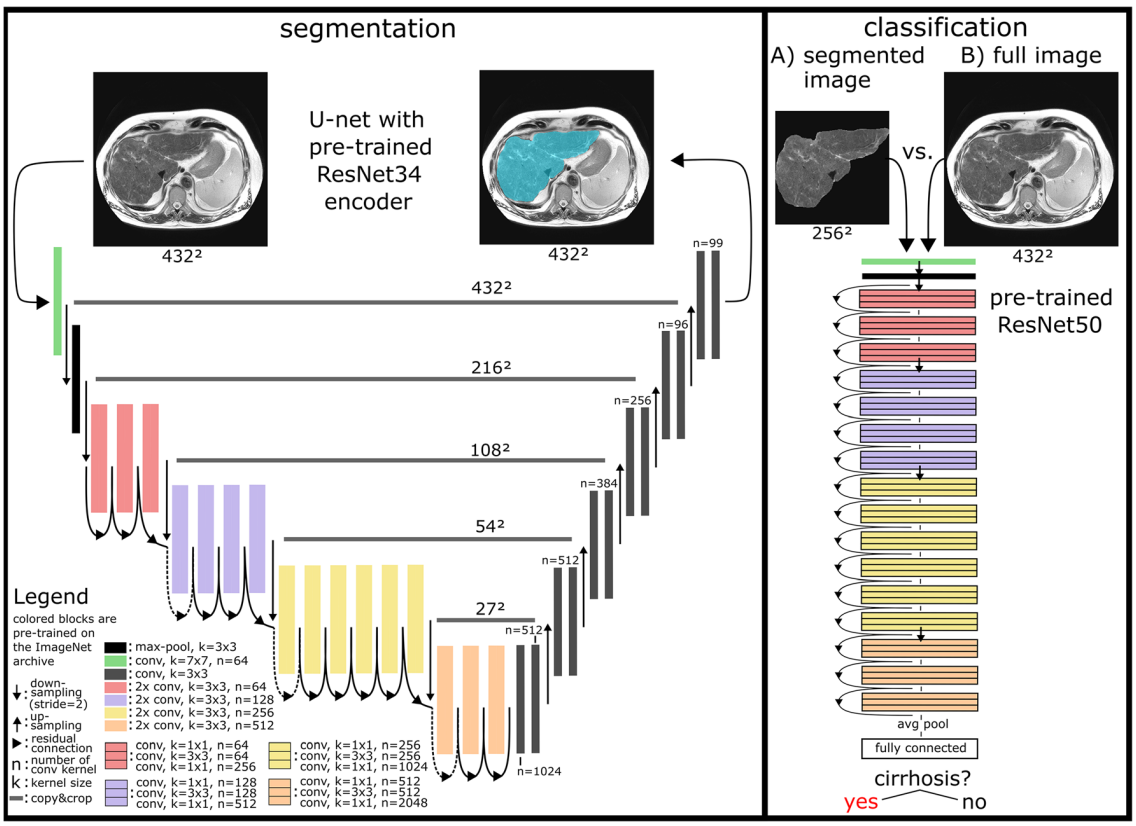


Fig. 2 Details of the presented deep transfer learning (DTL) pipeline for detection of liver cirrhosis. The segmentation network (left) is based on a U-net architecture, with a ResNet34 convolutional neural network (CNN) as encoder, pre-trained on the ImageNet archive. For the classification task (right), a pre-trained ResNet50 CNN was employed. The classification performance of the DTL pipeline including liver segmentation (**A**) was compared to a classification based on the original, unsegmented images (**B**)

task in both pipelines. The model was pre-trained on the ImageNet archive and implemented in pytorch's torchvision package [19]. Detailed descriptions of the segmentation and classification CNN architectures can be found in Fig. 2 and Supplement S4.

The DTL methods developed in this work were trained in two phases. First, only non-pretrained layers were trained and all pre-trained parameters of the convolutional layers were kept constant. To further investigate whether varying the pre-trained parameters may improve the reading performance of the CNN, the parameters of the pre-trained convolutional layers were made variable in a second phase. The one cycle learning rate policy was applied for fine-tuning of the pre-trained models for liver segmentation and classification of liver cirrhosis [20]. All experiments and evaluations were performed with python and fastai, a deep learning application programming interface for pytorch [21]. Further details of the experimental design and the hyper-parameters used for training are given in Supplement S5.

To compare the performance of the DTL analyses to the performance of healthcare professionals at different experience levels, validation and test data were also classified independently by a radiology resident (A.F.) with 4 years of experience in abdominal imaging and a board-certified radiologist (J.A.L.) with 8 years of experience in abdominal imaging.

The 95% confidence interval of the DTL-based classification accuracy was determined by the Clopper-Pearson method and a χ^2-test was performed to test for significant differences in accuracy between the DTL-based classification and the readers in SPSS Statistics 24 (IBM). For the test set, calculations of balanced accuracy, receiver operating characteristic, and precision-recall analyses were performed with scikit-learn 0.23.2 [22–24].

In order to assess the classification performance of the entire first pipeline (including prior segmentation), the segmentations of the CNN (instead of manual segmentations) were used for the validation and test set of the classification network. In addition to evaluating the method by its performance on the validation and test data set, gradient-weighted class activation maps (Grad-CAMs) were generated [25]. This technique is proposed to add visual information to radiological images, describing areas of the image that affect the prediction of the CNN [26]. These colored prediction maps were visually inspected and the image areas contributing to the CNN's prediction of cirrhosis were quantified separately for both patient groups.

Results

A total of 713 patients (342 female, mean age: 58 ± 14 years) were included. Of those, examinations of 572 patients were acquired at a field strength of 1.5 T. The remainder were examined on 3.0 T. A total of 553 patients (248 female, mean age: 60 ± 12 years) with a confirmed diagnosis of liver cirrhosis based on clinical or histopathological criteria were included (Fig. 1). The control group consisted of 160 subjects (94 female, mean age: 49 ± 18 years) without history of liver disease. A training set with 505 subjects (244 female, mean age: 58 ± 14 years), a validation set with 104 subjects (49 female, mean age: 57 ± 14 years), and a test set with 104 subjects (49 female, mean age: 58 ± 15 years) were compiled by random selection, while maintaining the proportion of control patients to patients with cirrhosis. The DTL method for segmentation of the liver in the transverse T2-weighted MRI images developed on the training set showed Dice values of 0.984 for the validation set and 0.983 for the test set.

In the subsequent training of the classification network ResNet50 for the identification of cirrhosis based on segmented images, an accuracy (ACC) of 0.99 (95% confidence interval: 0.95–1.00) for validation data (vACC) and 0.96 (0.90–0.99) for test data (tACC) was achieved. For the classification on unsegmented images, vACC was 0.97 (0.92–0.99) and tACC was 0.95 (0.89–0.98). The accuracy of the DTL pipeline for classification of cirrhosis with prior segmentation of the organ was significantly higher compared to the resident (vACC = 0.88, $p < 0.01$; tACC = 0.91, $p = 0.01$) as well as the board-certified radiologist (vACC = 0.96, $p < 0.01$; tACC = 0.90, $p < 0.01$) (Table 1). Modifications of pre-trained parameters did not improve segmentation and classification accuracy significantly (Table 2). On the test set, a balanced accuracy value of 0.90 was observed for the DTL method based on unsegmented images. Balanced accuracy values of 0.92 were observed for the DTL method based on segmented images, as well as for the radiology resident and board-certified radiologist. For the DTL method, the balanced accuracy of 0.92 is derived from a sensitivity of 1, which was higher than that of the radiology resident and board-certified radiologist (0.91, 0.89) and a specificity of 0.83, which was lower than that of the radiology resident and board-certified radiologist (0.92, 0.96).

Receiver operating characteristic and precision-recall curves for the test data set are shown in Fig. 3. For the DTL method trained on segmented images, an area under the curve (AUC) of 0.99 and an average precision (AP) of 0.97 and for the DTL method trained on the unsegmented images, an AUC of 0.95, and an AP of 0.93 were determined.

Figure 4 shows exemplary images from the test set with colored maps indicating areas which were particularly relevant for the decision of the classifier. The results of the visual inspection are presented in Table 3. In the first pipeline with upstream segmentation, the caudate lobe was highlighted in 47.5% of the images classified as cirrhosis and in 25% of the images classified as no cirrhosis. In every fifth (20.8%) of the segmented images classified as no cirrhosis, the transition zone of the caudate lobe to the image background was highlighted.

Table 1 Accuracy (ACC), balanced accuracy (BACC), sensitivity (Sens), and specificity (Spec) for identification of liver cirrhosis for validation (vACC, vBACC, vSens, vSpec) and test (tACC, tBACC, tSens, tSpec) of the deep transfer learning (DTL) method based on unsegmented images and based on images with prior segmentation of the liver. The accuracy of the DTL approaches was also compared to a radiological resident and a board-certified radiologist. Statistical difference was assessed by χ^2-test

Reader/method	vACC	*p* value (vAcc)	tACC	*p* value (tAcc)	vBACC	tBACC	vSens	tSens	vSpec	tSpec
ResNet50 (segmented liver)	0.99	-	0.96	-	0.99	0.92	0.99	1	1	0.83
ResNet50 (full image)	0.97	$p = 0.04$	0.95	$p = 0.61$	0.97	0.90	0.98	1	0.96	0.79
Board-certified radiologist	0.96	$p < 0.01$	0.90	$p < 0.01$	0.98	0.92	0.95	0.89	1	0.96
Radiology resident (4th year)	0.88	$p < 0.01$	0.91	$p = 0.01$	0.93	0.92	0.85	0.91	1	0.92

In the second pipeline, based on unsegmented images, additional highlighted areas outside of the liver were identified. In images classified as cirrhosis, the spleen area was highlighted in 6%, the stomach area in 22.5%, and the gastroesophageal junction in 12.5%. In 29.2% of the CNN's negative predictions, spinal musculature was highlighted.

Discussion

This proof-of-principle study demonstrates the feasibility of automatic detection of liver cirrhosis by DTL based on a standard T2-weighted MRI. The deep learning approach with prior segmentation of the liver provides classification accuracy at expert level.

To date, no other work has investigated the use of a DTL approach for the detection of liver cirrhosis in standard T2-weighted MRI sequences. There are recent studies based on gadoxetic acid–enhanced MRI imaging that classifies fibrotic pathologies of the liver by methods of deep learning and radiomics [27, 28]. However, these methods are trained from scratch and they require a manual definition of region of interests. In contrast to that, the method proposed in the current study does not require manual segmentation since the liver is segmented automatically with high precision.

Table 2 Dice values of the segmentation convolutional neural network (CNN) and classification accuracy of liver cirrhosis of the classification CNN at different stages of the training experiments. In the first stage of training the segmentation CNN, a Dice score of 0.9828 was achieved by optimizing the convolutional layers of the random-initialized decoder and remaining the parameters of the pre-trained ResNet34 encoder unchanged. In the following three stages that started from the model state of the previous stage, only minor improvements of 0.001 of the Dice score were achieved. In these stages, the convolutional layers of the pre-trained ResNet34 encoder were made variable, whereby the learning rate (LR) increased linearly from the first to the last layer of the CNN. In the first stage of training the classification CNN, an accuracy of 0.99 for the segmented images and 0.97 for the unsegmented images were achieved by optimizing the output layer of the ResNet50 CNN only. The following stages that started from the best previous model state did not lead to an improvement in accuracy and showed only minor improvements of the cross-entropy loss. Also in the last three stages, where the convolutional layers of the pre-trained ResNet50 were made variable with learning rates increased linearly from the first to the last layer of the CNN, no improvement in accuracy could be observed. Detailed descriptions of the training experiments can be found in Supplement S5

	Training stage	Epochs	Max LR last layer decoder	Max LR first layer encoder	Dice on validation set	
Segmentation network (U-net like with ResNet34 encoder)	1	80	0.001	Frozen	0.9828	
	2	40	0.0005	0.000005	No improvement	
	3	40	0.0005	0.00005	0.9837	
	4	40	0.0005	0.0005	0.9838	
	Training stage	**Epochs**	**Max LR output layer**	**Max LR first layer**	**Accuracy and cross-entropy loss (segmented image)**	**Accuracy and cross-entropy loss (full image)**
Classification network (ResNet50)	1	80	0.1	Frozen	0.99, 0.1452	0.97, 0.325
	2	40	0.01	Frozen	No improvement	0.97, 0.2151
	3	40	0.001	Frozen	No improvement	No improvement
	4	40	0.0001	0.000001	No improvement	0.97, 0.2025
	5	40	0.0001	0.00001	0.99, 0.1450	No improvement
	6	40	0.0001	0.0001	0.99, 0.1339	No improvement

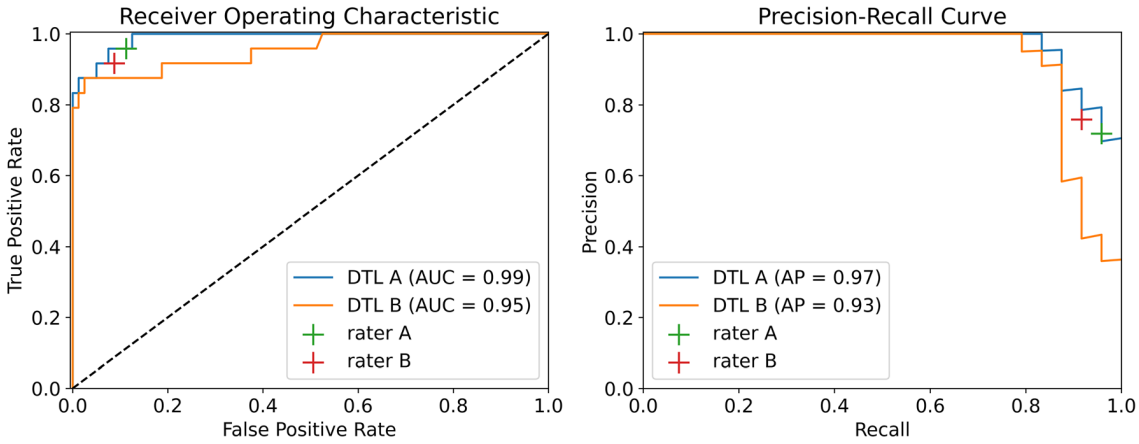


Fig. 3 Liver cirrhosis classification performance of the deep transfer learning (DTL) methods trained on the segmented images (DTL A) or unsegmented images (DTL B) and of the radiology resident (rater A) and the board-certified radiologist (rater B) on the test set, illustrated by receiver operating characteristic and precision-recall curves and area under the curve (AUC) and average precision (AP) values

Recent studies based on ultrasound imaging also used DTL methods pre-trained on the ImageNet archive [29, 30]. Of note, in both mentioned studies, the pre-trained parameters were not kept constant during training. Particularly the first few layers of the pre-trained CNNs have learned to recognize very general image features such as edges and shapes during the training with the ImageNet data set [31]. The ability to extract these features is a benefit of transfer learning, and therefore, other groups proposed to first optimize only the output layer of the network prior to changing the pre-trained parameters of the CNN [15, 32].

In order to examine whether altering the pre-trained parameters of the DTL methods is beneficial for the identification of cirrhosis, the CNNs were trained in two phases in this work, with frozen and unfrozen pre-trained parameters. Interestingly, the accuracy on the validation data set of both methods did not further increase by unfreezing the pre-trained parameters. Hence, the learned feature extraction capability from the training on the natural image data set of e.g. cars, animals, and buildings was generalized to identify liver cirrhosis on an expert level in standard T2-weighted MRI.

A further aim of our study was to investigate, whether prior segmentation of the liver is beneficial for this classification task. Interestingly, both variants (with and without prior segmentation) achieved high accuracy. However, the accuracy for the detection of liver cirrhosis was slightly higher for the DTL pipeline with prior segmentation. This result may be attributed to the following advantages of upstream segmentation:

i. The network is forced to focus on the area, where pathological alterations are primarily expected.
ii. Image areas that are not in focus of the analysis are prevented to have an impact on the normalization step [33].
iii. Using only the image areas of the organ allows to train the classification model with smaller image matrices and thus larger batch size, which is considered beneficial for the applied learning rate policy [20].

Table 3 Evaluation of the gradient-weighted class activation maps of the test set. The maps of the predictions of the deep transfer learning method, trained on segmented images and images without liver segmentation, were visually inspected and it was recorded which image areas were highlighted, separately for both patient groups. Note that several areas of the image were highlighted, so the percentages of the different image areas do not add up to 100% within a patient group. The liver areas were divided into left, right hepatic, and caudate lobe. For the segmented images, it was also noted whether image areas at the transition zone of the caudate lobe to the image background were highlighted. For the full images, highlighted areas near the stomach, spleen, gastroesophageal junction, and spinal muscles were observed

Unsegmented images	Patient group	Right hepatic	Left hepatic	Caudate lobe	Spleen	Stomach	Gastroesophageal junction	Spinal musculature
	Cirrhosis	53.8%	35%	22.5%	6.3%	22.5%	12.5%	2.5%
	No cirrhosis	83.3%	16.7%	0	0	8.3%	0	29.2%
Segmented images	Patient group	Right hepatic	Left hepatic	Caudate lobe	Border caudate lobe/-background		-	-
	Cirrhosis	53.8%	28.8%	47.5%	2.5%		-	-
	No cirrhosis	58.3%	20.8%	25%	20.8%		-	-

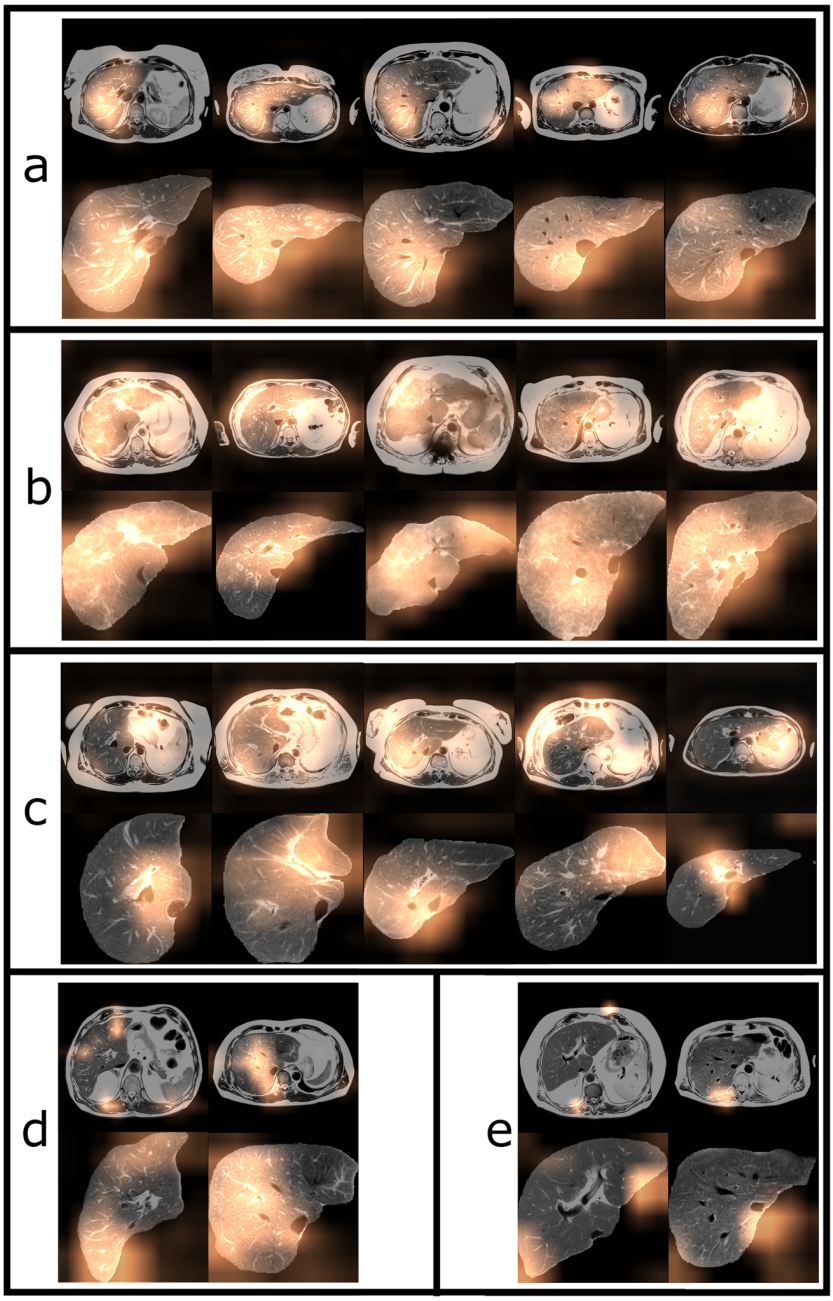


Fig. 4 Gradient-weighted class activation maps for unsegmented and segmented images from the test set. The overlays highlight regions that had high impact on classification in patients without cirrhosis (**a**) and patients with cirrhosis (**b**). Patients with and without cirrhosis that were correctly classified by the DTL methods but incorrectly classified by the certified radiologist are shown in **c**. Examples of images with a disagreeing classification of the two DTL methods, where the image was only correctly classified with prior liver segmentation are shown in **d**. Images that were misclassified by both DTL methods, but correctly classified by the certified radiologist are shown in **e**

For both methods, image areas relevant for the CNN's decision were investigated applying the Grad-CAM method [25]. The results indicate that the caudate lobe area is important for the DTL methods for the detection of liver cirrhosis trained on either segmented or unsegmented images. Interestingly, the Grad-CAM evaluations of the DTL method based on the unsegmented images showed that in some cases, image areas outside of the liver were relevant. This indicates that the CNN might also base the prediction of cirrhosis on accompanying signs of cirrhosis, such as spleen hypertrophy, venous alterations like fundus varices, or the general vital status of the patient according to muscle structure. This observation motivates further studies to investigate if deep learning methods may also reliably detect accompanying effects of cirrhosis.

Future work should also address whether a multi-task-learning architecture, which would simultaneously optimize segmentation and classification performance, has advantages over the presented pipeline. In addition, the method could be extended by an automated selection of the 2D slice at the level of the caudate lobe to allow fully automated prediction of cirrhosis based on T2-weighted imaging.

Our study has several limitations. First, the DTL model has been trained for the identification of liver cirrhosis only and

does not support the detection of very early signs of tissue fibrosis, which might be present in early hepatopathy. However, this was not the aim of this proof-of-principle study, but to investigate the hypothesis that ImageNet pre-trained models are generalizable to T2-weighted MRI imaging and allow the assessment of imaging features of liver cirrhosis. The investigation of an automated classification of early signs of tissue fibrosis and different stages of fibrosis will be the next step in the evaluation of deep transfer learning–based approaches based on standard T2-weighted MRI imaging.

Our study collective included a broad range of cirrhosis severities (according to the Child-Pugh score) and different etiologies of cirrhosis. To account for the difference in the number of patients with liver cirrhosis and patients without liver disease, additional performance measures were assessed. According to the balanced accuracy, the method trained on segmented images performs at expert level. However, the DTL method shows a higher sensitivity and a lower specificity compared to the board-certified radiologist, which may be a result of the class imbalance of the dataset. An expert level classification performance of the DTL method trained on segmented images is furthermore underlined by the precision-recall analysis.

Another limitation is that the classification was based solely on T2-weighted images. In contrast to that, additional pieces of information such as different MRI sequences as well as clinical and laboratory parameters are typically available for diagnosis in clinical routine. However, in our study, high diagnostic accuracy was shown for both the classifier and clinical experts, even if the diagnosis was based on only one anatomical sequence. Future studies may evaluate whether a multi-parametric approach or the inclusion of clinical parameters can further improve diagnostic performance.

Conclusion

This proof-of-principle study demonstrates the potential of DTL for the detection of cirrhosis based on standard T2-weighted MRI. The DTL pipeline for the image-based diagnosis of liver cirrhosis demonstrated classification accuracy at expert level. An application of the pipeline could support radiologists in the diagnosis of liver cirrhosis and has the potential to improve consistency of reading performance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-07858-1.

Funding Open Access funding enabled and organized by Projekt DEAL. The study was supported by a grant from the BONFOR research program of the University of Bonn (Application number 2020-2A-04). The funders had no influence on the conceptualization and design of the study, data analysis and data collection, preparation of the manuscript, and the decision to publish.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is PD Dr.med. Julian A. Luetkens.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval This retrospective study was approved by the institutional review board with a waiver of written informed consent.

Methodology
- Retrospective
- Diagnostic or prognostic study
- Performed at one institution

References

1. Volk ML, Tocco RS, Bazick J, Rakoski MO, Lok AS (2012) Hospital re-admissions among patients with decompensated cirrhosis. Am J Gastroenterol 107(2):247–252
2. Procopet B, Berzigotti A (2017) Diagnosis of cirrhosis and portal hypertension: imaging, non-invasive markers of fibrosis and liver biopsy. Gastroenterol Rep (Oxf) 5(2):79–89
3. Brown JJ, Naylor MJ, Yagan N (1997) Imaging of hepatic cirrhosis. Radiology 202(1):1–16
4. Rustogi R, Horowitz J, Harmath C et al (2012) Accuracy of MR elastography and anatomic MR imaging features in the diagnosis of severe hepatic fibrosis and cirrhosis. J Magn Reson Imaging 35(6): 1356–1364
5. House MJ, Bangma SJ, Thomas M et al (2015) Texture-based classification of liver fibrosis using MRI. J Magn Reson Imaging 41(2):322–328
6. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90

7. Nowak S, Faron A, Luetkens JA et al (2020) Fully automated segmentation of connective tissue compartments for CT-based body composition analysis: a deep learning approach. Invest Radiol 55(6):357–366
8. Zhu Y, Fahmy AS, Duan C, Nakamori S, Nezafat R (2020) Automated myocardial T2 and extracellular volume quantification in cardiac MRI using transfer learning–based myocardium segmentation. Radiol Artif Intell 2(1):e190034
9. Krogue JD, Cheng KV, Hwang KM et al (2020) Automatic hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell 2(2):e190023
10. Wang K, Mamidipalli A, Retson T et al (2019) Automated CT and MRI liver segmentation and biometry using a generalized convolutional neural network. Radiol Artif Intell 1(2):180022
11. Estrada S, Lu R, Conjeti S et al (2020) FatSegNet: A fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. Magn Reson Med 83(4):1471–1483
12. Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M (2020) FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. Neuroimage 219:117012
13. Kornblith S, Shlens J, Le QV (2019) Do Better ImageNet models transfer better? Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2661–2671
14. Shin H-C, Roth HR, Gao M et al (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298
15. Mormont R, Geurts P, Maree R (2018) Comparison of deep transfer learning strategies for digital pathology. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 2262–2271
16. Ravishankar H, Sudhakar P, Venkataramani R et al (2016) Understanding the mechanisms of deep transfer learning for medical images. In: Deep Learning and Data Labeling for Medical Applications. Springer, Cham, pp 188–196
17. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, pp 234–241
18. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
19. Paszke A, Gross S, Massa F et al (2019) PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, pp 8026–8037
20. Smith LN (2018) A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820
21. Howard J, Gugger S (2020) Fastai: a layered API for deep learning. Information 11(2):108
22. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
23. Brodersen KH, Ong CS, Stephan KE, Buhmann JM (2010) The balanced accuracy and its posterior distribution. The 20th International Conference on Pattern Recognition, IEEE, pp 3121–3124
24. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10(3):e0118432
25. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision, pp 618–626
26. Reyes M, Meier R, Pereira S et al (2020) On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol Artif Intell 2(3):e190043
27. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S (2018) Liver fibrosis: deep convolutional neural network for staging by using gadoxetic acid–enhanced hepatobiliary phase MR images. Radiology 287(1):146–155
28. Park HJ, Lee SS, Park B et al (2019) Radiomics analysis of gadoxetic acid–enhanced MRI for staging liver fibrosis. Radiology 290(2):380–387
29. Xue LY, Jiang ZY, Fu TT et al (2020) Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis. Eur Radiol 30:2973–2983
30. Lee JH, Joo I, Kang TW et al (2020) Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. Eur Radiol 30(2):1264–1273
31. Qin Z, Yu F, Liu C, Chen X (2018) How convolutional neural network see the world - a survey of convolutional neural network visualization methods. arXiv preprint arXiv:1804.11191
32. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer, Cham, pp 818–833
33. Collewet G, Strzelecki M, Mariette F (2004) Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging 22(1):81–91

European Radiology (2023) 33:884–892
https://doi.org/10.1007/s00330-022-09047-0

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Deep learning for standardized, MRI-based quantification of subcutaneous and subfascial tissue volume for patients with lipedema and lymphedema

Sebastian Nowak[1] · Andreas Henkel[1] · Maike Theis[1] · Julian Luetkens[1] · Sergej Geiger[1] · Alois M. Sprinkart[1] · Claus C. Pieper[1] · Ulrike I. Attenberger[1]

Received: 15 March 2022 / Revised: 20 June 2022 / Accepted: 21 July 2022 / Published online: 17 August 2022

Abstract

Objectives To contribute to a more in-depth assessment of shape, volume, and asymmetry of the lower extremities in patients with lipedema or lymphedema utilizing volume information from MR imaging.

Methods A deep learning (DL) pipeline was developed including (i) localization of anatomical landmarks (femoral heads, symphysis, knees, ankles) and (ii) quality-assured tissue segmentation to enable standardized quantification of subcutaneous (SCT) and subfascial tissue (SFT) volumes. The retrospectively derived dataset for method development consisted of 45 patients (42 female, 44.2 ± 14.8 years) who underwent clinical 3D DIXON MR-lymphangiography examinations of the lower extremities. Five-fold cross-validated training was performed on 16,573 axial slices from 40 patients and testing on 2187 axial slices from 5 patients. For landmark detection, two EfficientNet-B1 convolutional neural networks (CNNs) were applied in an ensemble. One determines the relative foot-head position of each axial slice with respect to the landmarks by regression, the other identifies all landmarks in coronal reconstructed slices using keypoint detection. After landmark detection, segmentation of SCT and SFT was performed on axial slices employing a U-Net architecture with EfficientNet-B1 as encoder. Finally, the determined landmarks were used for standardized analysis and visualization of tissue volume, distribution, and symmetry, independent of leg length, slice thickness, and patient position.

Results Excellent test results were observed for landmark detection (z-deviation = 4.5 ± 3.1 mm) and segmentation (Dice score: SCT = 0.989 ± 0.004, SFT = 0.994 ± 0.002).

Conclusions The proposed DL pipeline allows for standardized analysis of tissue volume and distribution and may assist in diagnosis of lipedema and lymphedema or monitoring of conservative and surgical treatments.

Key Points

- *Efficient use of volume information that MRI inherently provides can be extracted automatically by deep learning and enables in-depth assessment of tissue volumes in lipedema and lymphedema.*
- *The deep learning pipeline consisting of body part regression, keypoint detection, and quality-assured tissue segmentation provides detailed information about the volume, distribution, and asymmetry of lower extremity tissues, independent of leg length, slice thickness, and patient position.*

Keywords Deep learning · Magnetic resonance imaging · Lymphography · Leg · Subcutaneous tissue

Alois M. Sprinkart, Claus C. Pieper and Ulrike I. Attenberger contributed equally to this work.

✉ Alois M. Sprinkart
sprinkart@uni-bonn.de

[1] Department of Diagnostic and Interventional Radiology, Quantitative Imaging Lab Bonn (QILaB), University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

Abbreviations

CNN	Convolutional neural network
DL	Deep learning
MRL	MR-lymphangiography
SCT	Subcutaneous tissue
SFT	Subfascial tissue

European Radiology (2023) 33:884–892

Introduction

A chronic increase in leg circumference—either uni- or bilateral—can be caused by a range of pathological conditions: apart from venous disease and obesity, lymphedema and lipedema are recognized as major causes of increased extremity circumference [1, 2]. Lymphedema is characterized by soft tissue swelling caused by impaired lymphatic drainage leading to an accumulation of interstitial fluid. Through inflammatory reactions, a progressing deposition of subcutaneous fat, tissue fibrosis, and ultimately skin changes can be observed [2]. Lipedema is a disorder characterized by adipose tissue accumulation in the extremities and predominantly affects females. Patients typically present with a disproportionate distribution of body fat on the extremities despite a slender upper body and further symptoms such as fatigue and hyperalgesia. Additionally, lymphedema may develop in the affected patients [3, 4]. The pathophysiology of lipedema is so far poorly understood [1, 4, 5]. In patients suffering from either lipedema or lymphedema, both mechanic impairments—that can cause secondary arthritis or interfere with normal walking—and emotional disorders—resulting from an appearance that does not conform to today's ideal of beauty—can result in impaired quality of life [1].

Traditionally, the diagnosis of lipedema and lymphedema is made by clinical examination with evaluation of leg circumference, pitting edema, pain, typical clinical signs (e.g., Stemmer's sign), standardized anthropometric measurements (e.g., body weight, body mass index, waist-to-hip ratio, waist-to-height ratio), and patient history [1, 2, 4]. Especially since the introduction of microsurgical treatment options for lymphedema, a more in-depth evaluation of the affected legs by clinical MRI—e.g., as MR-lymphangiography (MRL)—has been introduced at specialized centers for treatment planning and therapy monitoring [6, 7]. In this respect, MRI is increasingly employed in the diagnosis, staging assessment, and follow-up of both lipedema and lymphedema and especially multi-echo T1-weighted images (e.g., using the DIXON technique) have been demonstrated to be useful for anatomical evaluation [8–11]. As simple anthropometric measures do not allow for separate assessment of subfascial and subcutaneous tissue and do not provide information on the volume distribution of these tissues along the entire extremities, it is therefore a logical step to leverage available imaging for precise volume assessment of these different compartments.

In recent years, DL methods have shown their potential to automate the quantification of tissue volumes in medical image analysis [12–15]. Therefore, DL could also provide a useful tool for automated imaging-based assessment of tissue volume in patients with suspected lipedema or lymphedema.

For a clinical application of artificial intelligence–based systems, it is important that the autonomous procedure has quality control mechanisms that are able to warn the treating physician in case of potentially limited validity of the measurement [14]. Quality control is not only important for evaluating individual examinations, but it can also be used to monitor the performance of the system over the time of deployment. The hardware requirements and the time required for inference are other aspects that affect the economics and accessibility of the automated systems, making a comparison of performance and efficiency of different DL models of interest.

Therefore, it was the aim of this study to develop a DL pipeline that allows to automatically extract precise normalized information of tissue volume, distribution, and symmetry from available MRI of the legs of patients with lipedema or lymphedema for standardized quantification of subcutaneous tissue (SCT) and subfascial tissue (SFT), while investigating the performance and efficiency of different architectures.

Material and methods

Dataset

This retrospective study was approved by the institutional review board with a waiver for written informed consent for data analysis. Consecutive patients who underwent clinical MRL examinations of the lower extremities between April 2016 and May 2017 were included into the study when they had either clinically diagnosed lymphedema (primary or secondary) or lipedema of the lower extremities. The indication for imaging was treatment planning (e.g., of lympho-venous anastomoses) in all patients. 3D DIXON MRL (slice thickness 5 mm, spacing between overlapping reconstructed slices 2.5 mm, in-plane resolution 1 mm) was performed as part of the pre-therapeutic diagnostic work-up on a 1.5-T MR system (Ingenia; Philips Healthcare) to assess gross and lymphatic anatomy as well as presence and extent of lymphatic run-off impairment. Clinical diagnosis was made by the referring experienced lymphologists based on the national guidelines for lymphedema and lipedema [4, 17].

Overall, 45 patients (42 female, mean age 44.2 ± 14.8 years) were examined during the selected time period and were included into the study. Of 45 patients, 36 (80%) suffered from lymphedema (13 primary, 23 secondary) and 9/45 (20%) from lipedema, with all men having secondary lymphedema and receiving MRL for treatment planning of lympho-venous anastomoses. Exclusively, DIXON water images were used for method development. In total, the dataset consisted of 18,760 slices in axial orientation. Data were randomly split into a training set for five-fold cross-validation of 40 (38 female, mean age 45.0 ± 15.5 years) cases and a hold-out test of 5 cases (4 female, mean age 37.4 ± 4.5 years) set. Detailed information on imaging parameters and image pre-processing prior to training can be found in Supplement S1.

The ground truth generation for the landmark detection was performed manually using Slicer 3D [18]. For the tissue segmentation, semi-manual tools and AI-assisted annotation were applied by a research assistant (S.N. with 3 years of experience in medical image segmentation). All annotations were finally approved by a board-certified radiologist (C.C.P. with 10 years of experience in lymphatic imaging). Further information on annotation can be found in Supplement S2.

The DL pipeline was finally also applied to four different use cases of routine clinical practice to demonstrate the clinical utility of the presented method for assessing tissue volume, distribution, and symmetry of the lower extremities and for monitoring of conservative or surgical treatment.

Method development for leg normalization

Figure 1 shows an overview of the developed pipeline consisting of two landmark detection methods and a quality-controlled tissue segmentation method.

(i) Leg model regression

In the first landmark detection method, a 2.5D CNN encoder determines the relative foot-head position of each axial slice within a standardized leg model by regression. To create the standardized leg model, the mean distances between the manually defined landmarks were determined for the entire dataset (ankle-knee: 95.0 ± 6.7 cm, knee-symphysis: 99.1 ± 6.1 cm, symphysis-femoral head: 16.7 ± 1.8 cm). The distances between the landmarks were normalized to the mean distance between ankles and knees, resulting in the relative positions −1, 0, 1.045, and 1.220 for ankles, knees, symphysis, and femoral heads within the leg model. The position values of the slices between the landmarks were linearly interpolated. Two numbers were then assigned to each axial slice, which corresponded to the relative position of that slice in the leg model for the left and for the right leg. Subsequently, image areas superior to the femoral heads were excluded from further analyses.

(ii) Keypoint detection

In the second landmark detection method, an additional 2.5D CNN encoder detects the image coordinates of the landmarks in coronal reconstructed slices using keypoint detection. To create the coronal reconstructed slices, the cropped images are down-sampled to an isotropic resolution of 2.5 mm. Then, slice by slice, the center of mass of the body mask was shifted in the anterior-posterior direction to the center of the image. Subsequently, the image matrix was cropped at a distance of 25 mm anterior and 25 mm posterior from the center of the image. This area contained all landmarks.

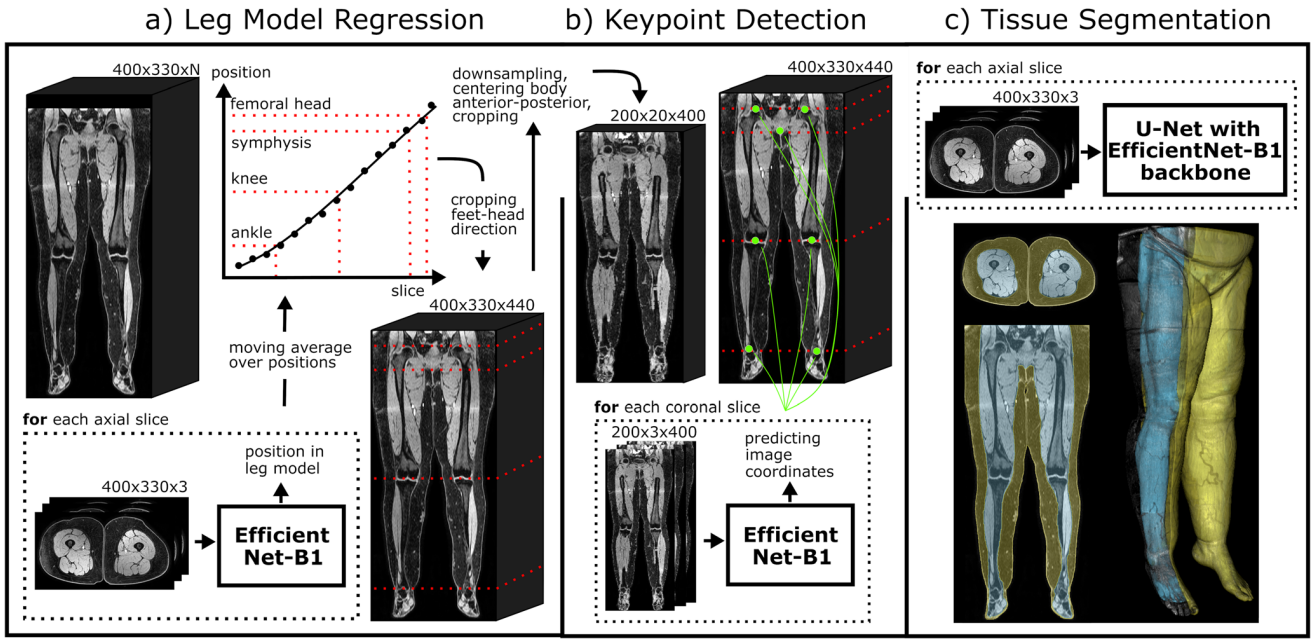


Fig. 1 Overview of the DL pipeline. (**a**) First, the 3D MRI scan is analyzed in axial slices by a 2.5D EfficientNet-B1 to identify the relative foot-head position of each slice with respect to a leg model consisting of ankles, knees, symphysis, and femoral heads. Afterwards, the image dataset is automatically cropped to the legs. (**b**) To increase the accuracy of the leg normalization, all landmarks are predicted by another 2.5D EfficientNet-B1 in coronal slices of a down-sampled cropped image using keypoint detection, where the lower limbs were centered slice-by-slice in anterior-posterior direction to the image center. (**c**) Then, a 2.5D U-Net with EfficientNet-B1 as backbone is used for segmentation of subcutaneous adipose tissue and subfascial tissue volume in axial slices. Finally, the identified landmarks and tissue volumes are combined to allow standardized quantification of the tissues (see Figs. 3 and 4)

For processing 3D information, 2.5D CNN encoders with three axial slices spaced 5 mm apart between each slice were used as input channels for all CNNs used in the current study. Also, different versions of a modern implementation of the established ResNet (ResNet18, ResNet34, ResNet50), as well as different versions of the recently introduced EfficientNets (B0, B1, B2, B3, B4), were implemented for both landmark detection methods [19, 20]. The most appropriate model was selected based on validation performance and model efficiency in terms of the number of trainable parameters and floating point operations required. For training the landmark detection methods, five-fold cross-validation was used and testing was performed with an ensemble of the cross-validated models.

Method development for tissue segmentation

Model

2.5D models with a U-Net-like architecture were investigated to segment the subcutaneous adipose tissue in the 3D MRI scans. Again, different versions of ResNet (ResNet18, ResNet34) and EfficientNet (B0, B1, B2, B3, B4) encoders were implemented for the U-Net model [19, 20]. The application of ResNet as an encoder of a U-Net has already been demonstrated to be able to segment the liver with high precision in MRI [21]. In addition, a CDFNet was trained, which was recently presented for abdominal adipose tissue segmentation in DIXON MRI images and computed tomography [13, 14]. Again, the most appropriate model was selected based on validation performance and model efficiency in terms of the number of trainable parameters and floating point operations required. Subsequently, the chosen 2.5D network architecture was also trained to perform segmentation on sagittal and coronal slices to investigate if a multi-view approach is beneficial for the current segmentation task [13].

As with the landmark detection methods, five-fold cross-validation was used for training. Testing was performed with an ensemble of the cross-validated models. Detailed information on the network architectures used as well as the hyperparameters used for training the tissue segmentation and landmark detection methods can be found in Appendix S3 and S4.

Quality control

For automatic assessment of segmentation quality, the entropy of the probability map of the segmentation models was used as a metric to predict the prediction uncertainty as in terms of the Dice score as proposed in previous studies [14, 16].

In the current study, two linear regression models were trained. One based on the entropy of the entire probability map of the 3D segmentation, as proposed in the original work, and another which considers the entropy slice by slice [16]. By this, it should be investigated whether this allows a local evaluation of the quality and thus represents a beneficial extension of the 3D approach. Only slices between the ankles and femoral heads with segmentations larger than 10% (12.7 cm^2) of the image section were considered. The linear regression models were trained with the predicted segmentations of all validation cases of the cross-validated tissue segmentation method and tested on the hold-out test set. Pearson correlation (r) coefficients were calculated with SciPy 1.6.3 [22].

Results

Leg normalization

EfficientNet-B1 was chosen as the most suitable model for both landmark detection methods used for leg normalization as it showed excellent performance in the five-fold cross-validation while having the least number of trainable parameters and floating point operations, resulting in a prediction time per patient of 2.7 s for the first method and 0.1 s for the second method on an NVIDIA Titan RTX graphics processing unit (GPU). Low mean deviations (Δz) between the predictions of the validation cases of the cross-validated landmark detection models and the manually defined ground truth were observed for the leg model regression (Δz = 6.6 ± 2.7 mm) and for the keypoint detection (Δz = 6.6 ± 3.2 mm). The mean sex-specific deviations were Δz-female = 6.4 ± 2.6 mm, Δz-male = 10.0 ± 4.4 mm for the leg model regression and Δz-female = 6.6 ± 3.2 mm, Δz-male = 7.9 ± 4.0 mm for the keypoint detection.

The ensemble of all cross-validated models showed also low mean deviations on the hold-out test set (leg model regression: Δz = 5.6 ± 5.6 mm; keypoint detection: Δz = 6.9 ± 4.4 mm). Considering an acceptable deviation of up to 10 mm in the test set, 85.7% of the landmarks detected by leg model regression and 74.3% of the landmarks detected by keypoint detection were correct. Using the predictions in an ensemble, the performance increased to Δz = 4.5 ± 3.1 (100% < 10 mm).

Tissue segmentation

EfficientNet-B1 was also chosen as the most suitable model for tissue segmentation as it showed again excellent segmentation performance on axial slices in the five-fold cross-validation with mean Dice scores of 0.982 ± 0.007 for segmenting SCT and of 0.989 ± 0.003 for segmenting SFT. The mean prediction time per patient was 8 s on an NVIDIA RTX 3090 GPU. Dice scores were consistently above 0.95 for both genders (Dice score female: SCT = 0.983 ± 0.007, SFT = 0.989 ± 0.003; Dice score male: SCT = 0.967 ± 0.005, SFT = 0.984 ± 0.003).

Combining the predictions of three multi-view models, each segmenting on either axial, coronal, and sagittal slices, did not improve the already very high segmentation performance (Dice SCT: 0.980 ± 0.008; Dice SFT: 0.987 ± 0.004).

An ensemble of the five-fold cross-validated EfficientNet-B1 models applied to axial slices achieved also excellent Dice scores on the test set (Dice SCT: 0.989 ± 0.004; Dice SFT: 0.994 ± 0.002).

Detailed information and illustrations on the model selection for tissue segmentation and landmark detection can be found in Supplement S5.

Quality control

Both linear regression models (based on 3D volumes and 2D slices) demonstrated a high correlation between the entropy of the subcutaneous tissue segmentation probability map and the segmentation quality represented by the Dice score (3D volumes: SCT $r = -0.76$ $p < 0.001$, SFT $r = -0.75$ $p < 0.001$; 2D slices: SCT $r = -0.78$ $p < 0.001$, SFT $r = -0.76$ $p < 0.001$). Low mean deviation between predicted and actual Dice score were observed when applying the models to the hold-out test set (3D volumes: ΔDice SCT: 0.003 ± 0.002; ΔDice SFT: 0.001 ± 0.001; 2D slices: ΔDice SCT: 0.003 ± 0.002; ΔDice SFT: 0.002 ± 0.003). Figure 2 shows the two regression models and also illustrates the application of the models for automatic identification of cases with lower segmentation quality.

Use cases

The trained DL pipeline was applied to different use cases of routine clinical practice to create leg normalized visualizations, which are shown in Figs. 3 and 4.

Discussion

This work presents a DL method for standardized quantification of subcutaneous and subfascial tissue of the lower extremities in patients with lipedema and lymphedema, which has the potential to provide an in-depth description of shape, volume, and asymmetry.

Modern imaging techniques have become increasingly important in the work-up of patients with suspected lipedema and lymphedema or lymphatic leakages [6, 7, 11, 23, 24]. Especially high-resolution 3D MRL has shown to be useful for planning of new surgical therapeutic options of lymphatic diseases [25] and may also be helpful in treatment follow-up. Usually morphological sequences are part of a MRL protocol and allow for structural assessment of the affected legs. Therefore, it is a logical consequence to apply the capabilities of DL to available morphological 3D imaging to automatically extract information about the exact tissue volumes that might be otherwise unused, which could lead to a more objective assessment of edematous diseases compared to conventional measurements.

Spatial standardization of identified tissues allows comparison between examinations independent of leg length, slice thickness, and position, ultimately allowing comparison of tissue volume distributions between initial and follow-up scans of a patient. To achieve automated standardized analysis, two tasks were solved by utilizing DL, namely tissue segmentation and landmark detection. As a further step towards clinical application, the proposed pipeline in the current study includes a segmentation quality control approach as proposed in a previous work [16]. As an extension to this method that based on the entire 3D volume, we additionally developed a linear regression model trained on each slice of the 3D scan. This allows to assess local quality of the segmentation process and is therefore more sensitive to local effects, e.g., caused by imaging artifacts.

For both the landmark detection and tissue segmentation methods, a 2.5D approach incorporating three slices was chosen. This approach has significantly lower computational costs compared to 3D CNNs, allowing analysis of the high-resolution MRI scans without prior down-sampling, while reducing hardware requirements and time needed for inference. Since excellent results were already observed for the 2.5D approach, the inclusion of more 3D related information through a multi-view approach was not found to be beneficial for the given tasks. Furthermore, the performance and efficiency of different CNN models for landmark detection and tissue segmentation were investigated in this work. The recently released EfficientNet, which showed state-of-the-art performance on the ImageNet dataset at the time of its release while maintaining very efficient computational requirements, was observed also to be high performant and efficient for medical landmark detection and tissue segmentation. Employing efficient models is also of interest for the use of DL in routine clinical practice, as they can reduce costs by further lowering hardware requirements and inference time.

A detailed comparison with previous work on the quantification of tissue volumes for lymphedema assessment in patients with breast cancer using manual landmark definition and non-DL algorithms, as well as previous work on body part detection in medical imaging, can be found in Supplement S6 [8, 26–28].

Our study has several limitations. First, MRI images from routine clinical practice of patients with lipedema and lymphedema, but no patients who are solely obese and have no edematous alterations, were used for the development of the method. However, we assume, although it was not explicitly tested in this study, that the deep learning pipeline also works in purely obese patients without edema, as the high Dice

European Radiology (2023) 33:884–892

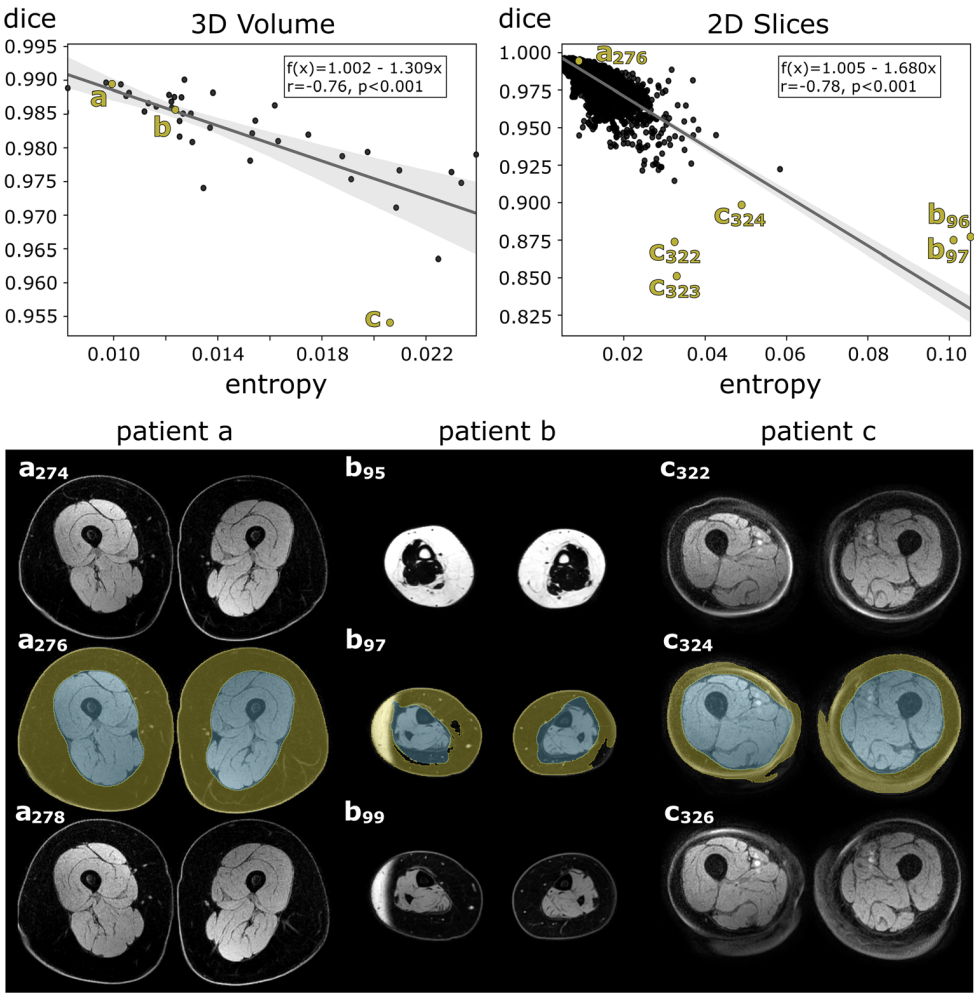


Fig. 2 The two linear regression models developed for quality control of the tissue segmentation convolutional neural network (CNN) are shown in the upper section of the figure. These are used for predicting the segmentation quality of the subcutaneous tissue class in terms of the Dice score. The first regression model was based on the entropy of the entire probability map of the 3D segmentation (top left). A second regression model was developed to predict segmentation quality slice by slice (top right). Gray areas represent 95% confidence intervals. Pearson correlation coefficient (*r*) along with the two-tailed p-value is given in the boxes. The lower section of the figure shows the 3 channel inputs of the 2.5D segmentation CNN for three patients (a, b, c), respectively, whose entropy of probability map and Dice scores are highlighted in the plot above. The digits represent the slice numbers. Excellent overall segmentation quality with high Dice scores and low entropy was observed for the majority of the entire 3D volumes and 2D slices (c.f. patient a). The slice-wise prediction of the Dice score allows to additionally capture local effects on segmentation quality caused, e.g., by water-fat swap (as seen in patient b) or partial volume artefacts (as seen in patient c). For patients b and c, adjacent artifact-affected slices, which also had low predicted Dice scores, are also highlighted in the plot above

values show that the tissue regions of the patients used for method development, where edema is not present, were also segmented with high precision. Also, data for method development were mainly from female patients. This is due to the fact that lymphedema as well as lipedema occurs predominantly in the female population and data from routine clinical practice was used for this study. However, excellent performance values for landmark detection and tissue segmentation were also observed in male patients of the validation sets. Furthermore, the deep learning method was developed using DIXON water images from a single MRI scanner only. Multicenter trials are warranted to proof the general applicability. The use of the algorithm will be enabled for collaborative multi-center studies on reasonable request (https://qilab.de). Also, at the current stage, there was no further investigation of the segmented tissues with respect to fluid infiltrations, which have implications for treatment strategy of lipedema and lymphedema. In this respect, the presented approach may be used as a basis for further quantitative analyses of tissue properties in future studies, e.g., by multi-parametric imaging. Lastly, the proposed method has not been evaluated for e.g. treatment response assessment in a clinical trial so far. However, we demonstrate potential use cases of the method showing examples for tissue volume assessment, evaluation of asymmetrical tissue proportions, and the evaluation of volume changes after surgical treatment. Future studies should

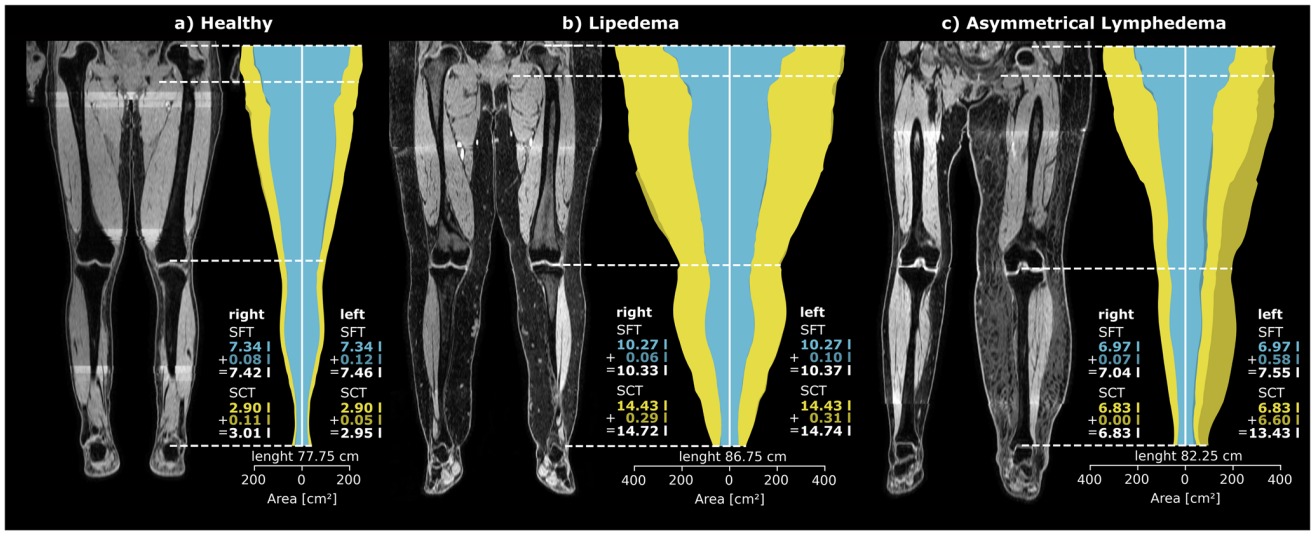


Fig. 3 Use cases for assessment of volume, distribution, and symmetry utilizing volume information from MRI. On the left (**a**) is a patient (female, 45 years old) without swelling of the lower extremities, in the middle (**b**) is a patient with lipedema (female, 46 years old), and on the right (**c**) is a patient with asymmetric left secondary lymphedema (female, 66 years old). Cumulative axial tissue areas are displayed per slice for each patient, with the distribution of the subfascial tissue (SFT) shown in blue and of the subcutaneous tissue (SCT) in yellow separated for the left and right leg between the femoral heads and the ankles. The detected landmarks are indicated by dotted lines. In order to highlight the differences in tissue volume between the two legs, asymmetric tissue portions are shown in darker blue for SFT and darker yellow for SCT. This is particularly apparent in the illustration of the patient with asymmetrical lymphedema (**c**). Next to the right and left leg, the tissue volumes are indicated in liters with corresponding colored font, and the total volume of SFT and SCT is indicated with white font

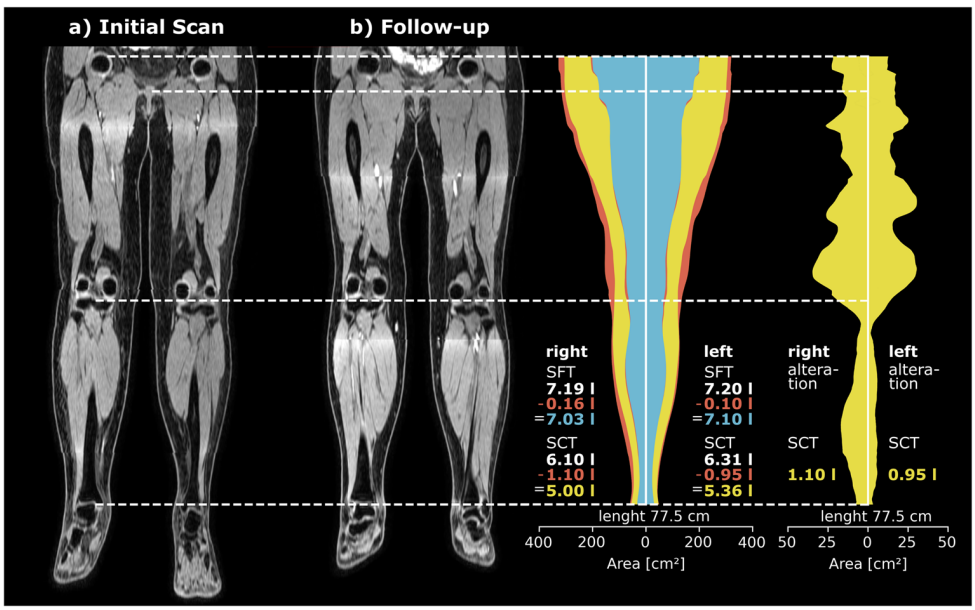


Fig. 4 Use case for evaluating success of surgical treatment. The figure illustrates normalized visualizations of a pre-therapeutic and 1-year follow-up scan of a lymphedema patient (female, 55 years old) who received surgical treatment (lympho-venous anastomoses). Cumulative axial tissue areas for the follow-up examination are illustrated, with the distribution of the subfascial tissue (SFT) shown in blue and of the subcutaneous tissue (SCT) in yellow separated for the left and the right leg between the femoral heads and the ankles. The differences in tissue volumes between the initial and the follow-up scan, i.e., tissue portions that have decreased in the course of the treatment, are indicated in red color. Next to the right and left leg, the total volume of SFT and SCT measured at the initial examination is indicated with white font, the total volume of SFT and SCT measured at the follow-up examination is indicated with blue and yellow font, and the decrease in volume is indicated with red font. On the right side of the figure, the alterations in SCT volume between initial and follow-up scan is presented in yellow and in a different scale to highlight where predominantly decrease of tissue volume has occurred during the course of treatment

evaluate the clinical value of the method for diagnosis, treatment planning, and treatment monitoring of lipedema and lymphedema against or in compliment of conventional anthropometric measurements.

Conclusion

This study presents a DL system for standardized and objective analysis of tissue volume, distribution, and symmetry based on MRI in patients with suspected lipedema or lymphedema.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-09047-0.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is PD Dr. med. Claus C. Pieper.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was waived by the institutional review board (University of Bonn).

Ethical approval This retrospective study was approved by the institutional review board.

Methodology
- retrospective
- experimental study
- performed at one institution

References

1. Kruppa P, Georgiou I, Biermann N, Prantl L, Klein-Weigel P, Ghods M (2020) Lipedema-pathogenesis, diagnosis, and treatment options. Dtsch Arztebl Int 117(22-23):396
2. Warren AG, Brorson H, Borud LJ, Slavin SA (2007) Lymphedema: a comprehensive review. Ann Plast Surg 59(4):464–472
3. Halk AB, Damstra RJ (2017) First Dutch guidelines on lipedema using the international classification of functioning, disability and health. Phlebology 32(3):152–159
4. Reich-Schupke S, Schmeller W, Brauer WJ et al (2017) S1 guidelines: lipedema. J Dtsch Dermatol Ges 15(7):758–767
5. Forner-Cordero I, Szolnoky G, Forner-Cordero A, Kemény L (2012) Lipedema: an overview of its clinical manifestations, diagnosis and treatment of the disproportional fatty deposition syndrome–systematic review. Clin Obes 2(3-4):86–95
6. Lohrmann C, Foeldi E, Langer M (2009) MR imaging of the lymphatic system in patients with lipedema and lipo-lymphedema. Microvasc Res 77(3):335–339
7. Pieper CC (2020) Nodal and pedal MR lymphangiography of the central lymphatic system: techniques and applications. Semin Intervent Radiol 37(3):250–262
8. Borri M, Gordon KD, Hughes JC et al (2017) Magnetic resonance imaging–based assessment of breast cancer–related lymphoedema tissue composition. Invest Radiol 52(9):554
9. Arrivé L, Derhy S, Dahan B et al (2018) Primary lower limb lymphoedema: classification with non-contrast MR lymphography. Eur Radiol 28(1):291–300
10. Cellina M, Martinenghi C, Panzeri M et al (2021) Noncontrast MR lymphography in secondary lower limb lymphedema. J Magn Reson Imaging 53(2):458–466
11. Cellina M, Gibelli D, Soresina M et al (2020) Non-contrast MR lymphography of lipedema of the lower extremities. Magn Reson Imaging 71:115–124
12. Koitka S, Kroll L, Malamutmann E et al (2021) Fully automated body composition analysis in routine CT imaging using 3D semantic segmentation convolutional neural networks. Eur Radiol 31: 1795–1804
13. Estrada S, Lu R, Conjeti S et al (2020) FatSegNet: a fully automated deep learning pipeline for adipose tissue segmentation on abdominal DIXON MRI. Magn Reson Med 83(4):1471–1483
14. Nowak S, Theis M, Wichtmann BD et al (2021) End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT. Eur Radiol. https://doi.org/10.1007/s00330-021-08313-x
15. Nowak S, Faron A, Luetkens JA et al (2020) Fully automated segmentation of connective tissue compartments for CT-based body composition analysis: a deep learning approach. Invest Radiol 55(6):357–366
16. Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T (2020) Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans Med Imag 39: 3868–3878
17. Wilting J, Bartkowski R, Baumeister R et al (2017) S2k Leitlinie Diagnostik und Therapie der Lymphödeme. AWMF online. Availiable via https://www.awmf.org/uploads/tx_szleitlinien/-0011_S2k_Diagnostik_und_Therapie_der_Lymphoedeme_2019-07.pdf. Accessed 6 Jan 2022
18. Fedorov A, Beichel R, Kalpathy-Cramer J et al (2012) 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 30(9):1323–1341
19. He T, Zhang Z, Zhang H, Zhang Z, Xie J, Li M (2019) Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 558-567

20. Tan M & Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning, pp. 6105-6114
21. Nowak S, Mesropyan N, Faron A et al (2021) Detection of liver cirrhosis in standard T2-weighted MRI using deep transfer learning. Eur Radiol 31:8807–8815
22. Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17(3):261–272
23. Pieper CC, Feisst A, Schild HH (2020) Contrast-enhanced interstitial transpedal MR lymphangiography for thoracic chylous effusions. Radiology 295(2):458–466
24. Pieper CC, Hur S, Sommer CM et al (2019) Back to the future: lipiodol in lymphography-from diagnostics to theranostics. Invest Radiol 54(9):600–615
25. Forte AJ, Boczar D, Huayllani MT et al (2021) Use of magnetic resonance imaging lymphangiography for preoperative planning in lymphedema surgery: a systematic review. Microsurgery 41(4): 384–390
26. Yan Z, Zhan Y, Peng Z et al (2016) Multi-instance deep learning: discover discriminative local anatomies for bodypart recognition. IEEE Trans Med Imaging 35(5):1332–1343
27. Zhang P, Wang F, Zheng Y (2017) Self supervised deep representation learning for fine-grained body part recognition. In IEEE 14th International Symposium on Biomedical Imaging, pp. 578-582
28. Yan K, Lu L, Summers RM (2018) Unsupervised body part regression via spatially self-ordering convolutional neural networks. In 2018 IEEE 15th International Symposium on Biomedical Imaging, pp. 1022-1025

4. Discussion with references

The publications in this work demonstrate the potential of AI-based pipelines for extracting and analysing quantifiable features from routine clinical radiological images. In addition to DL-based automated tissue segmentation, anatomical landmark identification, or disease detection, the works highlight techniques that can be used to design pipelines that offer potential utility in clinical practice.

In the first publication, an automated end-to-end pipeline for body composition analysis in CT is presented, which includes automatic anatomical landmark detection, tissue segmentation and quality control. For autonomous systems operating on routine clinical data, mechanisms that warn of potentially invalid analyses, e.g., due to artefacts, are of great importance. For body composition analysis, beam hardening artefacts caused by metal implants in the spine can make the assessment of the surrounding muscle infeasible. Here, it was shown that Machine Learning and conventional image processing methods can be used to overcome this problem by assessing segmentation quality or detecting artefacts. A second challenge for autonomous systems is the handling of variable scan lengths, resolutions and the resulting variable image sizes. The connections of a CNN are defined according to a given input size, so that in principle a CNN can only analyse image matrices of this fixed size. In this work, a patch-wise CNN, where the image is split into parts of predefined size prior to analysis, is proposed as the first instance of the pipeline offering a possible solution to this challenge. The results of this work can be considered as further advancement of an earlier study on tissue segmentation for body composition analysis, which was also published during the period of this thesis (Nowak et al., 2020).

In the second paper, a DL pipeline is proposed for the detection of liver cirrhosis in clinical MRI. In this work, the utility of transfer learning was demonstrated by showing that the image features extracted from a frozen CNN previously trained with over a million natural images of, e.g., cars, animals, and buildings, could be used for expert-level disease detection. This result is of particular interest for the development of DL algorithms based on routine clinical data, where for a cohort of patients the number of images can be limited, especially when addressing rarer

diseases. Furthermore, it has been shown that a limitation of the DL analysis to the segmented organs is beneficial compared to analysis of the entire image. Finally, gradient-weighted class activation, a method aiming for explainable AI, were also used to generate saliency maps that provided insights on relevant liver regions. However, conclusions about individual decisions are difficult to draw from the coarse highlighted areas, which is a criticized limitation of these methods (Ghassemi et al., 2021). This questions the utility of these methods for detecting invalid analyses and thus their use as potential quality control instances of a DL pipeline in routine clinical practice. The findings of this work formed the basis of another work on characterization of an alcohol-related etiology of liver cirrhosis, which was also published during the period of this thesis (Luetkens et al., 2022).

The third paper presents a pipeline for standardized assessment and visualization of leg tissue distribution in patients with lip- and lymphedema in MRI. This work demonstrated how DL can extract otherwise unused precise tissue volume information of the entire leg from imaging in a standardized way, allowing comparisons between examinations for monitoring disease progression or treatment success. Conventionally, tissue volume alterations in these patients are assessed at the body surface using anthropometric measurements, such as leg circumferences, and not by determining exact volumes from imaging. Therefore, this study is an example of how DL has the potential not only to automate tedious analyses, but also to improve diagnostic and monitoring procedures. However, this remains to be proven in future clinical studies of this pipeline. Also, the novel regression of the relative position of each axial slice within a standardized leg model not only improves the accuracy of landmark detection in the two-stage approach, but also serves as input to the DL pipeline that can handle variable scan lengths. Furthermore, it was shown that a slice-by-slice application of a previously presented method for quality control has advantages and allows for identification of local artefacts that influence segmentation quality.

The AI-pipelines developed in this thesis have following limitations. One limitation is that the training with supervised learning required time-consuming manual or semi-manual annotation of data. This included the annotation of tissues of interest, landmark identification, or especially the retrospective compilation of patient

cohorts from the radiological and clinical data systems. Since radiological reports or letters of the referring clinicians are commonly in free-text format, the retrospective identification of a patient cohort with a disease of interest requires opening and reading these texts. Using all available data from a very large clinical data system for supervised learning is often hindered by the need for time-consuming manual curation. Therefore, there is a great need for unlocking radiological and clinical databases for the development of AI methods. To achieve this, unsupervised learning methods that do not require manually annotated data are in demand. In recent years, text-based transformer models, which can be pre-trained by unsupervised learning techniques, have emerged as a state-of-the-art language processing architectures (Brown et al., 2020). An ongoing study of our institute is investigating how text-based transformer models can be efficiently developed on-site by also employing unsupervised learning techniques to retrospectively categorize and thereby unlock radiological report databases for data-driven medicine. In addition, recent developments in multimodal analysis of image-text pairs are of great interest for radiological application. These techniques are currently used to create AI-generated synthetic images based on text inputs and are trained in two phases. In the first phase, an image encoder and a text encoder are jointly trained to both generate similar feature representations for similar or corresponding image-text pairs. In the second phase, the abstract feature representations of the text encoder are used as a seed for an image generating model (Ramesh et al., 2022). Although the generation of synthetic medical images is not expected to be of great clinical utility, leveraging the abstract feature representation of a medical image for AI-based generation of radiology reports is of great interest (Hosny et al., 2018). The potential of these modern techniques for use in radiology will be investigated in future studies.

Another limitation is that the AI methods of this work, which were trained by supervised learning on annotated data, are exclusively task-specific and can only be applied to the one specific operation for which they were developed. The inability to perform more than one task is a typical limitation of current AI-tools and described as Artificial Narrow Intelligence (ANI) (Hosny et al., 2018; Shevlin et al., 2019). AI systems that, like humans, have the ability to apply their cognitive

resources to a variety of different tasks are referred to as Artificial General Intelligence (AGI) and to date have not been accomplished (Shevlin et al., 2019). The recent and promising developments of text-based transformer architectures and the results achieved by massive scaling of data, hardware resources and trainable parameters of these models indicate that the path to AGI may be feasible in the future (Brown et al., 2020). This progress towards AGI by scaling transformers is also reflected in a recent work that developed a transformer that could be trained to handle over 600 different tasks from various domains, such as conducting dialogues, outperforming humans playing Atari games, operating a robot arm, and many more (Reed et al., 2022). Among other benefits, the authors suggest that generic models have advantages over domain-specific solutions due to more efficient use of computation and the increase in variety and amount of training data by incorporating numerous tasks. Future work could explore the potential of these concepts to overcome ANI of current radiological AI-tools.

In conclusion, DL-based QIA pipelines for three different scenarios of clinical routine were presented that enable disease detection, extraction of standardized and objectively quantifiable image features, or monitoring of disease progression or therapy success. These systems demonstrate how challenges of routine clinical data can be overcome and thus analysis pipelines can be designed that have the potential to provide true benefits for routine clinical practice.

References

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. Advances in neural information processing systems 2020; 33: 1877-1901

Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021; 3(11): e745-e750

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018; 18(8): 500-510

Luetkens JA, Nowak S, Mesropyan N, Block W, Praktiknjo M, Chang J, Bauckhage C, Sifa R, Sprinkart AM, Faron A, Attenberger UI. Deep learning supports the differentiation of alcoholic and other-than-alcoholic cirrhosis based on MRI. Sci Rep 2022; 12(1):1-8

Nowak S, Faron A, Luetkens JA, Geißler HL, Praktiknjo M, Block W, Thomas D, Sprinkart AM. Fully automated segmentation of connective tissue compartments for CT-based body composition analysis: a deep learning approach. Invest Radiol 2020, 55(6): 357-366

Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with clip latents. arXiv preprint 2022; arXiv:2204.06125

Reed S, Zolna K, Parisotto E, Colmenarejo SG, Novikov A, Barth-Maron G, Gimenez M, Sulsky Y, Kay J, Springenberg JT, Eccles T, Bruce J, Razavi A, Edwards A, Heess N, Chen Y, Hadsell R, Vinyals O, Bordbar M, de Freitas N. A generalist agent. arXiv preprint 2022; arXiv:2205.06175

Shevlin H, Vold K, Crosby M, Halina M. The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge. EMBO reports 2019; 20(10): e49177

5. Acknowledgements

I would like to express my deepest gratitude to Univ.-Prof. Dr. med. Ulrike I. Attenberger, who enabled me to conduct this thesis under her supervision at the Department of Diagnostic and Interventional Radiology and who was always supportive during this time. I would also like to express my deepest gratitude to Prof. Dr. rer. nat. Jürgen Hesser, who guided the work in right directions with his extraordinary professional expertise. I am extremely grateful for the trust and support of the other members of the thesis committee, Univ.-Prof. Dr. med. Manuel Ritter and also Priv.-Doz. Dr. rer. nat. Wolfgang Block, who was always available and supportive in personal concerns.

Also, I want to express my deepest appreciation especially to Priv.-Doz. Dr.-Ing. Alois Martin Sprinkart, who has been a mentor to me since supervising my Bachelor thesis, investing in me and advocating for me. Many thanks also to my colleague Maike Theis, with whom I had good cooperation, but also a lot of fun during the time of writing this thesis. Also many thanks to my other colleagues and co-authors at the UKB.

I want to thank my parents, who have encouraged me throughout my life, supported me in good and bad times and thus played a major role in my personal and professional development. I would like to thank my big brother, who has always been and will always be a role model for me. Lastly, I would like to thank my girlfriend, who supported me during the period of this thesis and patiently endured one or the other fanatical raving about current Deep Learning methods during covid lockdown evenings.

List of journal publications

08/18 *Nowak S* & Sprinkart AM. Synchronization and Alignment of Follow-up Examinations: a Practical and Educational Approach Using the DICOM Reference Coordinate System. J Digit Imaging 2019; 32(1): 68-74

03/20 *Nowak S*, Faron A, Luetkens JA, Geißler HL, Praktiknjo M, Block W, Thomas D, Sprinkart AM. Fully automated segmentation of connective tissue compartments for body composition analysis: a Deep Learning Approach. Invest Radiol 2020; 55(6): 357-366

07/20 Faron A, Sprinkart AM, Kuetting DLR, Feisst A, Isaak A, Endler C, Chang J, *Nowak S*, Block W, Thomas D, Attenberger UI, Luetkens JA. Body composition analysis using CT and MRI: Intra-individual intermodal comparison of muscle mass and myosteatosis. Sci Rep 2020; 10: 11765

05/21 *Nowak S*, Mesropyan N, Faron A, Block W, Reuter M, Attenberger UI, Luetkens JA, Sprinkart AM. Detection of liver cirrhosis in standard T2-weighted MRI using deep transfer learning. Europ Radiol 2021; 31(11): 8807-8815

09/21 *Nowak S*, Theis M, Wichtmann BD, Faron A, Froelich MF, Tollens F, Geißler HL, Block W, Luetkens JA, Attenberger UI, Sprinkart AM. End-to-end automated body composition analyses with integrated quality control for opportunistic assessment of sarcopenia in CT. Europ Radiol 2021; 32(5): 3142-3151

12/21 Faron A, Opheys NS, *Nowak S*, Sprinkart AM, Isaak A, Theis M, Mesropyan N, Endler C, Sirokay J, Pieper CC, Kuetting D, Attenberger UI, Landsberg J, Luetkens JA. Deep Learning-Based Body Composition Analysis Predicts Outcome in Melanoma Patients Treated with Immune Checkpoint Inhibitors. Diagnostics 2021; 11(12): 2314

05/22 Luetkens JA, *Nowak S*, Mesropyan N, Block W, Praktiknjo M, Chang J, Bauckhage C, Sifa R, Sprinkart AM, Faron A, Attenberger UI. Deep learning supports the differentiation of alcoholic and other-than-alcoholic cirrhosis based on MRI. Sci Rep 2022; 12: 8297

08/22 *Nowak S*, Henkel A, Theis M, Luetkens JA, Geiger S, Sprinkart AM, Pieper CC, Attenberger UI. Deep learning for standardized, MRI-based quantification of subcutaneous and subfascial tissue volume for patients with lipedema and lymphedema. Europ Radiol 2022; https://doi.org/10.1007/s00330-022-09047-0

Peer review activity

2021 Review for “European Radiology” (twice)

2022 Review for “Die Radiologie”

Speaker at conferences

12/20 Mesropyan N, *Nowak S*, Faron A, Theis M, Block W, Reuter M, Attenberger UI, Luetkens JA, Sprinkart AM, Detection of liver cirrhosis in standard T2-weighted MRI images using Deep Transfer Learning, RSNA 2020 - 106. Annual Meeting of the Radiological Society of North America

03/21 *Nowak S*, Faron A, Mesropyan N, Reuter M, Block W, Kütting DLR, Attenberger UI, Luetkens JA, Sprinkart AM. Deep learning for differentiation of liver cirrhosis aetiology from clinical MRI. ECR 2021 Book of Abstracts. Insights Imaging (2021); 12: 75

05/21 *Nowak S*, Mesropyan N, Faron A, Theis M, Block W, Reuter M, Attenberger UI, Luetkens JA, Sprinkart AM. Detektion einer Leberzirrhose in der T2-gewichteten MRT mittels Deep-Transfer-Learning. Rofo 2021; 193(1): 7

07/22 *Nowak S*, Henkel A, Theis M, Block W, Attenberger UI, Pieper CC, Sprinkart AM. Deep-learning for standardized, MRI-based quantification of subcutaneous and subfascial tissue volume of the lower extremities for patients with lip- and lymphedema. ECR 2022