Aus der

Radiologischen Universitätsklinik
Abteilung Diagnostische und Interventionelle Radiologie

Validierung der automatischen Organsegmentierung von Leber und Milz im Ganzkörper MRT im Rahmen der Nationalen Kohorte

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Medizin

der Medizinischen Fakultät
der Eberhard Karls Universität
zu Tübingen

vorgelegt von
Fetzer, Lukas Michael
2025

Dekan:	Professor Dr. B. Pichler
1. Berichterstatter:	Professor Dr. S. Gatidis
2. Berichterstatter:	Professor Dr. R. Bares
Tag der Disputation:	09.10.2025

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

NAKO	Nationale Kohorte
MRT	Magnetresonanztomographie
KORA	Kooperative Gesundheitsforschung in der Region Augsburg
KI	Künstliche Intelligenz
CNN	Convolutional Neural Network
MA	Multi-Atlas
CF	Classification Forest
CPU	Central Processing Unit
GPU	Graphic Processing Unit
NN	Neural Network
ML	Machine Learning
DCNN	Deep Convolutional Neural Networks
VIBE	Volume Interpolated Breathhold Examination
DICOM	Digital Imaging and Communications in Medicine
MITK	Medical Imaging Interaction Toolkit
BMI	Body Mass Index
SD	Standardabweichung

Validierung der automatischen Organsegmentierung von Leber und Milz im Ganzkörper MRT im Rahmen der Nationalen Kohorte

1. Einleitung

1.1 Gesundheitsstudie Nationale Kohorte

Im Bereich der medizinischen Forschung gibt es die Bereiche Primär- und Sekundärforschung. In der Primärforschung werden neue Daten gewonnen, zum Beispiel mithilfe von Studien. In der Sekundärforschung gelangt man mittels bereits vorhandener Daten zu neuen Analysen oder Ergebnissen. In den Bereich der Primärforschung fallen wiederum epidemiologische sowie klinische Forschungen als auch die Grundlagenforschung (1-3)

Bei epidemiologischen Studien kann weiterhin in Interventions- und Beobachtungsstudien unterschieden werden. Ein Beispiel für eine Interventionsstudie wäre die Randomisiert-kontrollierte-Studie. Beobachtungsstudien umfassen passive patienten- oder probandenbezogene Datenerhebungen und lassen sich wiederum in Fall-Kontroll-Studien, Querschnittsstudien und Kohortenstudien einteilen.

Kohortenstudien haben das Ziel, mögliche Zusammenhänge zwischen Auslösern und auftretenden Krankheiten aufzudecken. Sie können pro- oder retrospektiv gestaltet sein. Im prospektiven Design wird ein Kollektiv aus Probanden erstellt und oft über einen langen Zeitraum beobachtet, befragt und untersucht. Mögliche auslösende Umstände für Erkrankungen, und ob diese gar auftreten, sind dabei unbekannt und stehen im Fokus der Detektion. Bei retrospektiven Kohortenstudien hingegen wird auf bereits erhobene Daten zurückgegriffen. Das bedeutet, dass der Unterschied der beiden Verfahren im Zeitpunkt der Datenakquise liegt (3). Auf diese Weise gelingt es Risiken und mögliche Auslöser für bestimmte Erkrankungen abzuschätzen (3).

Die Nationale Kohorte (NAKO) ist eine über 25 bis 30 Jahre geplante Kohortenstudie in Deutschland. Die NAKO Studie zielt darauf ab, Daten über die Gesundheit der Probanden zu erheben und die Entstehung von Volkskrankheiten wie zum Beispiel kardiovaskulären Erkrankungen, Diabetes Mellitus, Krebs, Atemwegserkrankungen, neurodegenerativen und psychiatrischen Erkrankungen, muskuloskelettalen Erkrankungen oder Infektionen besser zu verstehen und deren Vorstufen besser erkennen zu können (4).

Organisiert wird diese Studie von verschiedenen Trägern, wie der Leibniz-Gesellschaft, der Helmholtz-Gemeinschaft, verschiedenen Universitäten und weiteren Einrichtungen in ganz Deutschland. Beabsichtigt ist eine Teilnehmerzahl von insgesamt 200.000 Personen im Alter von 20 bis 69 Jahren. Das Kollektiv der Studie wird geographisch repräsentativ in 18 verschiedenen Studienzentren in Deutschland rekrutiert. Dabei werden städtische, industrialisierte und auch ländliche Bereiche abgedeckt.

Fünf der Studienzentren sind mit einheitlichen Magnetresonanztomographen (MRT) bestückt und für die Ganzkörper-Magnetresonanztomographie-Bildgebung (Ganzkörper-MRT-Bildgebung) ausgestattet (siehe Abbildung 1)(4).

Alle Teilnehmer unterziehen sich einer Reihe von Untersuchungen. Der Umfang der durchgeführten Untersuchungen wird in drei Leveln unterschieden. Die als Level 1 bezeichnete Untersuchung wird bei allen Teilnehmern durchgeführt und umfasst ausführliche Anamnesegespräche, Fragebögen, körperliche Untersuchungen und labortechnische Untersuchungen. Diese labortechnischen Untersuchungen umfassen unter anderem Blut-, Speichel-, Urin- und Stuhlproben (4).

Bei etwa 40.000 randomisiert ausgewählten Teilnehmern werden weitreichendere Untersuchungen durchgeführt. Dieses zusätzliche Untersuchungsprogramm wird als Level 2 bezeichnet. Wiederum 30.000 Teilnehmer der Probanden von Level 2 erhalten als Level 3 Untersuchung eine Ganzkörper-MRT-Bildgebung, die zerebrale -, kardiale -, Körperfettgewebe -, thorakoabdominelle - und muskuloskelettale Bildgebung umfasst. Der zeitliche Abstand zwischen Level 1 und Level 3 Untersuchung soll den maximalen Abstand von zwölf Wochen nicht überschreiten (5). Repräsentativ werden die MRT Aufnahmen geografisch verteilt akquiriert (siehe Abbildung 1).

Sämtliche Teilnehmer sollen nach vier bis fünf Jahren erneut zu den Level 1 und Level 2 Untersuchungen geladen werden und dann im Abstand von zwei bis drei Jahren per Fragebogen zu Änderungen im Lebensstil oder gesundheitlichem Status befragt werden (4, 5).

Enorme Datenmengen von Blutanalysen, Anamnesen und Verhaltensweisen, Risikofaktoren und molekulare Untersuchungen können mithilfe der zusätzlichen MRT-Untersuchungen zu phänotypischen Profilen zusammengestellt werden. Die Analyse

dieser Datenmenge und -vielfalt ermöglicht Rückschlüsse auf gegenseitige Einflussnahmen von festgestellten Befunden und bedeutet die Gewinnung neuer Erkenntnisse über die Entstehung von Krankheiten in der Bevölkerung (3, 6, 7). Gerade bei Krankheiten mit multifaktorieller Genese erhofft man sich neue Einsichten, welche in Bezug auf die Studiendauer und die Menge an wiederholt durchgeführten Untersuchungen erkannt werden können (3, 6, 7).

Eine der größten Herausforderungen ist dabei die Auswertung der durch MRT gewonnenen Bilddatensätze. Eine einzelne MRT-Untersuchung besteht aus Hunderten von einzelnen Bildern und kann nur durch spezialisiertes Personal ausgewertet werden. Eine automatisierte Auswertung wäre nicht nur für die NAKO-Studie, sondern auch alle zukünftigen Untersuchungen ein Meilenstein.

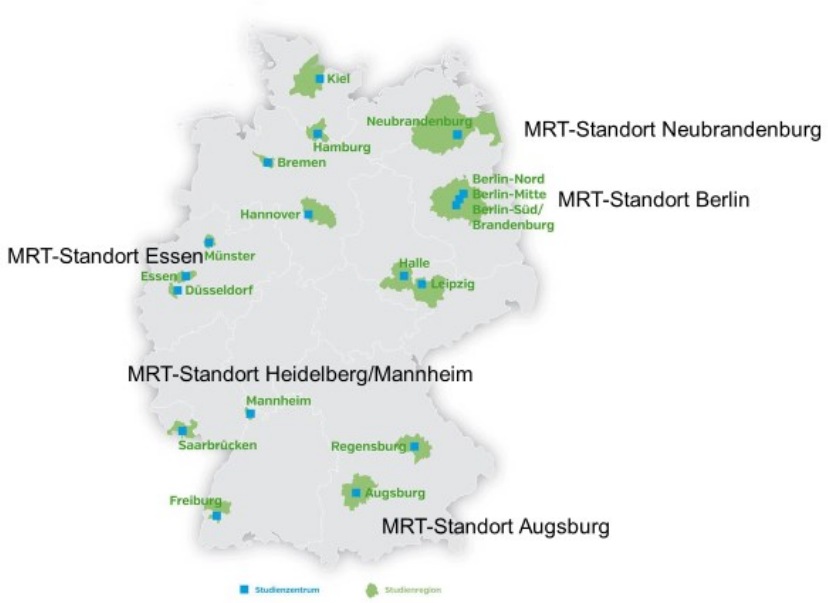

Abbildung 1, Studienzentren und MRT-Standorte der NAKO Studie, (https://nako.de/studienteilnehmer/studienzentren/) (8)

1.2 Medizinischen Bildgebung

1.2.1 Grundsätzliches

In der Medizin bedient man sich verschiedensten Bildgebungsverfahren. Es gibt mehrere etablierte Verfahren zur Bildgebung in der Medizin. Die Sonographie, welche Ultraschall zur Darstellung benutzt, ist zeitintensiv und benutzerabhängig. Dies bedeutet Schwierigkeiten in der Reproduzierbarkeit und eine eingeschränkte

Nutzbarkeit bei großangelegten Kohortenstudien (7, 9). Einige Verfahren wie das Röntgen oder die Computertomographie (CT) setzen dabei auf ionisierende Strahlung, welche Nebenwirkungen verursacht, und zur Entstehung von Erkrankungen führen kann (10). Die Magnetresonanztomographie kommt ohne den Einsatz der ionisierenden Strahlung aus und ist für bestimmte Fragestellungen eine schonendere Alternative zur CT (10). Mittels komplexer Magnetfeldtechnologie können ohne Einsatz ionisierender Strahlung Gewebe und anatomische Strukturen hochauflösend sichtbar gemacht werden (10). Dieses Verfahren gehört wie die CT Untersuchung zu den sogenannten Schnittbildverfahren. Durch Steigerung der Effizienz in Bereichen wie Aufnahmedauer und Betriebskosten kann die MRT-Untersuchung heute auch im großen Rahmen, wie oben beschrieben bei der NAKO, eingesetzt werden. Gerade für Studien mit großen Teilnehmerzahlen ist die MRT-Bildgebung hilfreich, da keine negativen Nebeneffekte oder eine wesentliche Beeinflussung des Körperzustandes auftreten (9).

Generell können bei MRT-Untersuchungen verschiedene Kontraste erzeugt werden, z.B. die sogenannte T1- und T2-Wichtung. Dabei beziehen sich die Unterschiede hauptsächlich auf den unterschiedlich erzeugten Weichteilkontrast. In der T1-Wichtung tritt fettreiches Gewebe und Kontrastmittel hell, sprich hyperintens oder signalstark, auf. Wasser, Knochen, Bänder, Sehnen und Luft dunkel, hypointens oder signalarm. In der T2-Wichtung stellt sich Wasser und Fettgewebe hyperintens und Knochen, Bänder, Sehnen und Luft hypointens dar (11).

Eine Herausforderung stellt die Handhabe mit Zufallsbefunden dar. Bei einer großen Anzahl zu untersuchenden Probanden werden zwangsläufig Zufallsbefunde erhoben, welche eventuell noch keine körperlichen Beschwerden hervorgerufen haben. Um individuell die Auswirkungen auf den jeweiligen Probanden abschätzen zu können, und um zu entscheiden, ob es sich um mitteilungsbedürftige Erkenntnisse handelt, wurde eine Expertengruppe innerhalb der NAKO gegründet, welche sich genau mit dieser Thematik beschäftigt(5).

1.2.2 Bildanalyse

Bei der medizinischen Bildanalyse folgt als erster Schritt die visuelle Begutachtung der erstellten Aufnahmen durch einen Radiologen.

Die Bildanalyse lässt sich in qualitative und quantitative Aspekte unterteilen. Diese umfassen verschiedene Teilbereiche wie zum Beispiel die Visualisierung, die Registrierung und die Segmentierung (12). Generell sind diese Prozesse untereinander verknüpft und bauen zum Teil aufeinander auf (siehe Abbildung 2)(12). Beispielsweise können die Bilddaten direkt visualisiert werden, oder erst registriert oder segmentiert werden. Segmentierte und registrierte Bilddaten können visualisiert werden, zur Quantifizierung ist allerdings oft eine vorherige Segmentierung notwendig.

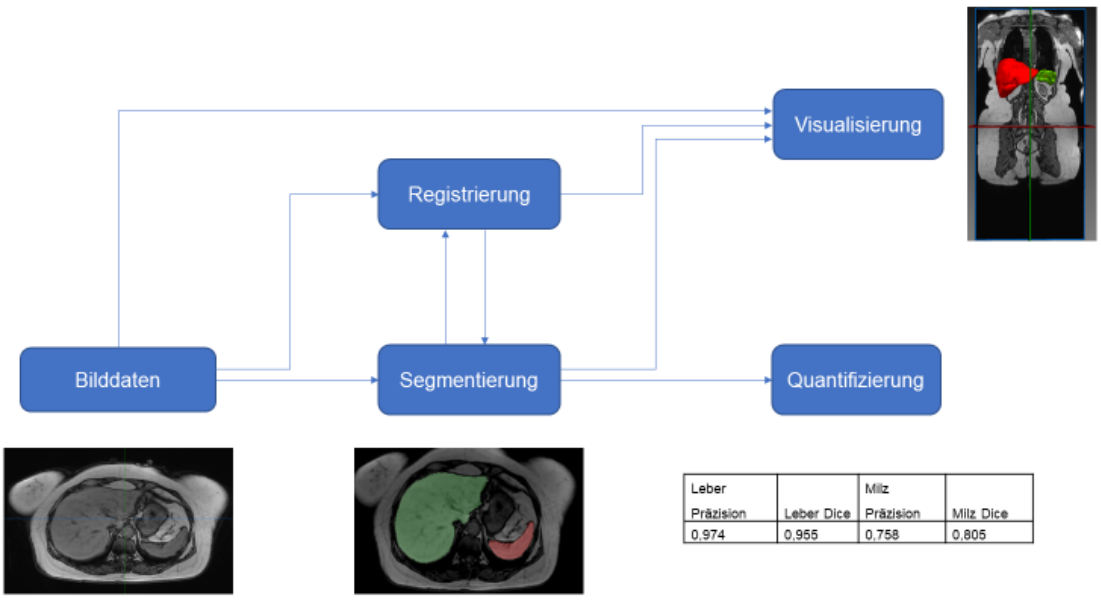

Abbildung 2, Schematische Darstellung der Abläufe in der medizinischen Bildanalyse mit Darstellung von Beispielen (5)

Ein wichtiger Schritt der Bildanalyse ist die Segmentierung (12). Bei der Segmentierung handelt es sich um ein Verfahren der digitalen Bildverarbeitung, Es werden zusammenhängende anatomische Strukturen (z.B. Organe) im Bilddatensatz identifiziert und lokalisiert und von nicht dazugehörenden Strukturen wie dem Hintergrund abgegrenzt. Dies bedeutet, dass alle zu der jeweiligen Struktur gehörenden zweidimensionalen Bildpunkte (Pixel) oder dreidimensionalen Gitterpunkte (Voxel) markiert werden. Die Segmentierung dient als Basis für weiterführende Verarbeitungen und kann für die Registrierung und Visualisierung eine Rolle spielen (12). Die Registrierung wird unabdingbar, sobald mehrere Bildaufnahmen, oder Aufnahmevarianten, sowie verschiedene

Untersuchungszeitpunkte in die Informationen mit einfließen. Ziel der Registrierung ist es unter anderem verschiedene Bilder in größtmöglicher Übereinstimmung anzugleichen, um Aufnahmen aus verschiedenen Modalitäten übereinander abzubilden. Bei der Visualisierung werden die Ergebnisse aus Segmentierung und Registrierung ausgewertet und Varianten zur eigentlichen Analyse angewandt, die Betrachtung einer MRT Aufnahme kann in zwei- und dreidimensionaler Ansicht erfolgen. In der quantitativen Bildanalyse werden dann verschiedenste Diameter ausgemessen, Volumina errechnet oder Signalstärken ermittelt (13).

Die Segmentierung von Bildgebungsdaten kann durch Zielvolumensegmentierung in der Bestrahlungsplanung ein wichtiger Schritt sein. Dort werden die Gewebe den zu therapierenden Pathologien in der Bildgebung segmentiert und dienen als Zielführung für die Bestrahlung (14).

Des Weiteren spielt die Lebersegmentierung eine Rolle bei der Volumetrie vor einer Lebendspende, um eine bessere Übersicht im Resektionsgebiet zu erhalten und die erforderliche *Livervolume to body size ratio* zu ermitteln (15).

Die Segmentierung von Organen in der medizinischen Bildgebung stellt eine Herausforderung dar. Diese manuelle Arbeit ist mit hohem Zeitaufwand verbunden und bindet Arbeitskräfte, es entstehen Aufwand und Kosten. Durch die Automatisierung dieses Prozesses fiele die aufwendige manuelle Bearbeitung und Segmentierung weg. Aufgrund des hohen Aufwands der manuellen Segmentierung wird versucht, diesen Prozess zu automatisieren.

Es wurden in der Vergangenheit verschiedenste Möglichkeiten zur automatisierten Organsegmentierung implementiert. Die klassische Bildverarbeitung, welche auf schwellenwertbasierter oder kantenbasierter Segmentierung fußt, war eines der ersten Verfahren, die versuchten eine Automatisierung in diesen Prozess einzubringen (16-18). Eine weitere Option ist das *Machine Learning* (ML) (18). Beispiele solcher ML Verfahren sind sogenannte *Classification Forests* (CFs), *Multi-Atlas Konzepte* (MA) oder *Convolutional Neural Networks* (CNN). In der vorliegenden Arbeit wird CNN als Verfahren gewählt, da es sich gegenüber CFs oder MA vor allem im Punkt Genauigkeit durchgesetzt hat (19). Ein weiterer Grund ist die Vorerfahrung mit dem CNN Algorithmus in der Kohortenstudie Kooperative Gesundheitsforschung in der Region Augsburg (KORA), wo bereits ein Algorithmus für ein ähnliches Unterfangen verwendet wurde (20).

1.3 Machine Learning

1.3.1 Grundsätzliches

Die immer weiter voranschreitende Entwicklung im Bereich der künstlichen Intelligenz (KI) bringt unter anderem auch Algorithmen hervor, welche automatisiert lernen, sich anpassen und sich fortlaufend optimieren können (21). Besonders in der oben beschrieben automatischen Bildanalyse großer zugrunde liegender Datenmengen rücken diese Algorithmen in den Fokus (21). ML beschreibt dabei den Prozess um die künstliche Erlernung von Wissen.

Es kann zwischen *unsupervised learning* und *supervised learning* im Bereich des ML differenziert werden (21). Beim *unsupervised learning* wird darauf abgezielt, dass der Algorithmus selbstständig zum Beispiel grafische Unterschiede in Form von Kontrastunterschieden erkennt, ohne dass vorher eine Markierung in Trainingsdaten vorgenommen wurde (21). Beispielsweise sollen so zusammenhängende Untergruppen in vorhandenen Daten identifiziert werden, (21). Im Falle des *supervised learning* werden dem Algorithmus klar vordefinierte Daten dargeboten, beispielsweise segmentierte Trainingsdaten, aus denen dann das vorgegebene Ziel erlernt werden soll (21).

Durch die Trainingsdaten werden Fragen und zugehörige Antworten vorgegeben, anhand derer der Lernalgorithmus eine allgemeine Lösung des gestellten Problems erlernen soll. Ein Beispiel für eine ML Anwendung ist die automatische Segmentierung von Organen auf medizinischen Bilddaten. Die Aufgabe besteht darin, jene Bildareale zu erkennen, die das Zielorgan zeigen; in entsprechenden Trainingsdaten sind diese Areale bereits z.B. von einem Menschen markiert. Der Algorithmus wird mittels *supervised learning* dann anhand der Trainingsdaten so optimiert, dass die gestellte Aufgabe möglichst gut auch auf neuen Daten außerhalb des Trainingsdatensatzes ausgeführt werden kann (21).

In Abbildung 3 ist dieser Ablauf schematisch dargestellt. Man erkennt die manuell segmentierten Aufnahmen, welche dem Algorithmus als Training dargeboten werden. Der Algorithmus erstellt ein Modell, welches dann zur automatischen Organsegmentierung benutzt werden kann.

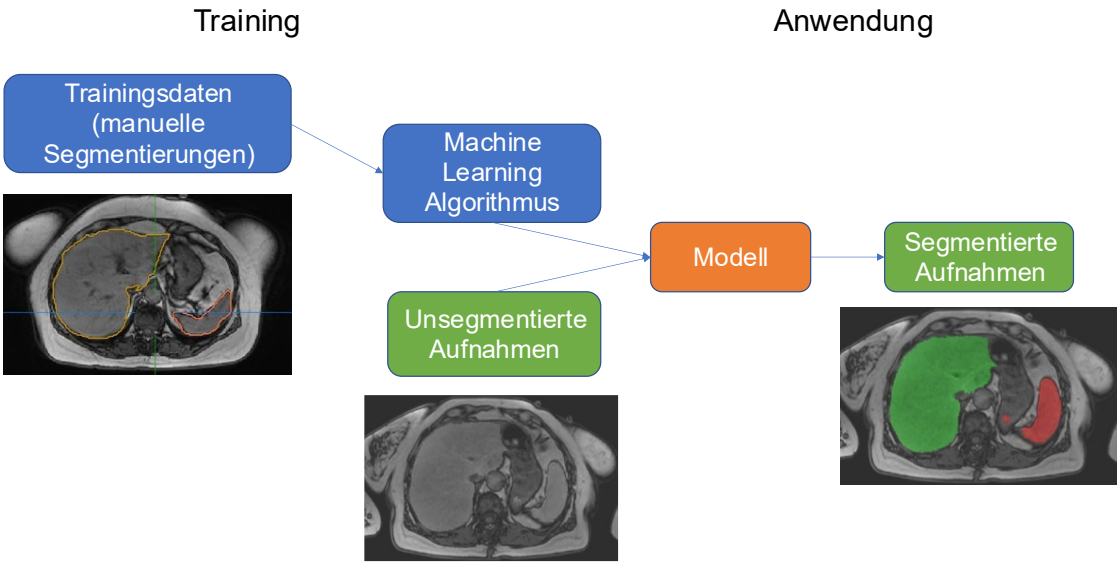

Abbildung 3, Schematische Darstellung des auf supervised learning *aufbauenden CNN Algorithmus. Trainingsdaten werden benutzt, um den Algorithmus zu trainieren. Dieser erstellt ein Modell, welches dann zur automatischen Segmentierung benutzt wird.*

1.3.2 Convolutional Neural Networks und Deep Learning

CNNs bilden tiefe Netzwerke aus mehreren Schichten. Sie werden besonders im Bereich der *computer vision*, zu Deutsch etwa Maschinelles Sehen, eingesetzt und gehören zur Klasse der *Neural Networks* (NN) (22). NN sind Modelle des ML im Bereich des *supervised* oder *unsupervised learning* Sie bestehen aus mehreren Ebenen, den sogenannten *layers*; aufgrund der großen Anzahl der *layers* (also der *Tiefe* der Netzwerke) spricht man auch von *deep learning* (DL) (21).

Zahlreiche Aufgaben wurden bereits mit CNN in Vorarbeiten bearbeitet, beispielsweise die automatische Bildklassifizierung (23, 24). Auch in der Segmentierung medizinischer Bildgebung kamen CNN bereits zum Einsatz (22, 25-27). Zu immer größerer Beliebtheit dieser Methode führten mehrere technologische Fortschritte. Unter anderem die Verfügbarkeit der notwendigen leistungsstarken Hardwarekomponenten wie *Graphic Processing Units* (GPU) und die Entwicklung der *backpropagation*, zu Deutsch Fehlerrückführung, welche effizient die Berechnung des Grades des Trainingsfehlers erlaubt. Das hohe Maß an Flexibilität der NN, die anspruchsvolle Verbindungen zwischen Trainingsdaten und Resultat zulassen (21), spielt ebenfalls eine Rolle. Die Flexibilität erlaubt zwar herausfordernde Modelle zu berechnen, birgt jedoch auch die Gefahr, dass bei geringer Menge an Trainingsdaten

der NN Algorithmus die Ausgangsdaten sozusagen auswendig lernt, anstatt neue Muster zu erlernen (sog. *overfitting*) (21). Daher ist eine passende Anzahl an Trainingsdaten und eine ausführliche Validierung wichtig, um diesem sogenannten *overfitting* zu begegnen.

NN beinhalten zwischen *input*- und *outputlayer* noch zusätzliche verborgene *layers*. Das Besondere an CNN ist, dass diese *convolutional* und *pooling layers* enthalten. Innerhalb des *convolutional layer* sind Neurone in zwei aufeinander folgenden *layers* nur miteinander verbunden, wenn diese räumlich gesehen nahe zusammenstehen. Im *pooling layer* werden die Wertigkeiten von Neuronen in räumlich angrenzenden Örtlichkeiten abgeschätzt (21). Diese Verarbeitung räumlicher Informationen ist besonders in der Bildanalyse entscheidend, da hier die räumliche Verteilung der Voxel den Bildinhalt prägt (28). Durch eine größere Anzahl an *layers* entstehen komplexere Versionen von CNN, sogenannte *deep convolutional neural networks* (DCNN) (29) .

Grundvoraussetzung dieser automatisierten Verfahren sind im Falle des *supervised learning* manuell segmentierte Trainingsdatensätze, die in die Algorithmen eingespeist werden. Anhand dieser Trainingsdatensätze kann der CNN Algorithmus adaptiv lernen. Um dies zu erreichen werden sogenannte Filter über das zu erlernende Bild bewegt. Stark vereinfacht beschrieben, werden rasterartig einzelne Abschnitte des Inputbildes abgetastet, stößt der Filter auf eine gesuchte Übereinstimmung, so markiert er die Stelle der Übereinstimmung im Outputbild (30). Das bedeutet, dass Filter wiederholt über einzelne Ausschnitte eines zu untersuchenden Objektes bewegt werden, um letztendlich optimierte Filtereinstellungen zu implementieren, mit denen z.B. automatisch Bilder analysiert werden sollen.

1.3.3 Performance des Machine Learning Verfahrens anhand von *learning curves*

Um die Performance eines ML Verfahrens zu ermitteln, können sogenannte *learning curves* hinzugezogen werden. Sie zeigen die Zeit gegenüber des Lernprogresses auf. Des Weiteren steht die Zeit für die *Erfahrung* des Algorithmus und der Lernprogress für die Verbesserung des Algorithmus. Eine hier in den Beispielen gewählte minimierende Metrik auf der Y-Achse zeigt, je kleiner der erreichte Wert, desto kleiner der Trainingsfehler. Erreicht die Kurve den Wert 0,0 auf der Y-Achse, würde dies für ein perfektes Lernen ohne Fehler sprechen. Sie können auch als *loss curves* bezeichnet werden, aufgrund der *loss*-Titulierung repräsentativ für die Fehlerrate. Je

nach Trainings- oder Validierungskurve kann dies als *training*- oder *validation loss* bezeichnet werden.

Anhand des Aussehens dieser Kurven können Fortschritte im Trainings- und Validierungsprozess festgestellt werden. Des Weiteren kann auch gezeigt werden, ob das Verfahren über genügend Kapazität verfügt, um die Datensätze zu verarbeiten, oder ob die Trainingsdauer passend gewählt ist. Dies kann dazu beitragen, je nach Aussehen der *learning curve* entsprechende Verbesserungen vorzunehmen.

Aus der Form von *learning curves* können verschiedene Informationen gewonnen werden, nämlich ob es sich um ein *underfit*, ein *overfit* oder ein *good fit* handelt, ob der Algorithmus mit ausreichender Kapazität für die Bearbeitung ausgestattet ist und ob die vorgesehene Trainingszeit passend ist (31). Beim *underfit* ist der zu trainierende Algorithmus nicht in der Lage die Aufgabe zu bewältigen und kann keine niedrige Fehlerrate erzielen (32). Dies kann sich durch eine generell flache Kurve zeigen, die Kurve des *validation loss* sollte gemeinsam mit dem *training loss* regressiv abfallen und sich auf einem möglichst niedrigen Niveau stabilisieren, zeigt hier jedoch einen linearen Verlauf mit großem Abstand zur anderen Kurve (31)(siehe Abbildung 4, Kurve 1). *Overfitting* kann auftreten, wenn der Algorithmus sich zu sehr auf die Trainingsdaten spezialisiert. Dadurch kann es zu einem Anstieg des *training*- oder *validation loss* im Laufe des Prozesses kommen, der Algorithmus erkennt neue Daten schwieriger und lernt ungewünschte Informationen oder Fehler aktiv mit. Es tritt vor allem bei zu langer Trainingszeit, zu hoher Kapazität für die gestellte Aufgabe und zu hoher Flexibilität des Algorithmus auf (31). In der *learning curve* verzeichnet der *training loss* eine Verminderung des erzielten Wertes auf der Y-Achse, und der *validation loss* sinkt zu einem gewissen Punkt ab und steigt dann wieder an (31)(siehe Abbildung 4, Kurve 2). Ein gewünschtes *good fit* in der *learning curve* tritt auf, wenn sich der *training loss* auf ein gewisses Level absinkt und sich stabilisiert und der *validation loss* sich ähnlich verhält (31) (siehe Abbildung 4, Kurve 3). Weiterführendes Training kann dann wiederum zum *overfit* führen (31). Der *training loss* ist meist niedriger als der *validation loss* und wird allgemein als *generalization gap* bezeichnet. Die *generalization gap* zeigt den Unterschied zwischen der Performance des Algorithmus bei den Trainingsdaten im Bezug zur Performance bei noch ungesehenen Daten aus gleicher Quelle (33). Außerdem können sich die Kurven so darstellen, dass der *training loss*, also die Kurve für die Fehlerrate oder –verminderung während des

Trainingsprogresses, zunächst langsam abfällt und bis zum Ende des Durchganges ein Minimum erreicht (31)(siehe Abbildung 4, Kurve 4). Dies kann auf eine ausreichende Kapazität und eine unzureichende Trainingsdauer hinweisen (31).

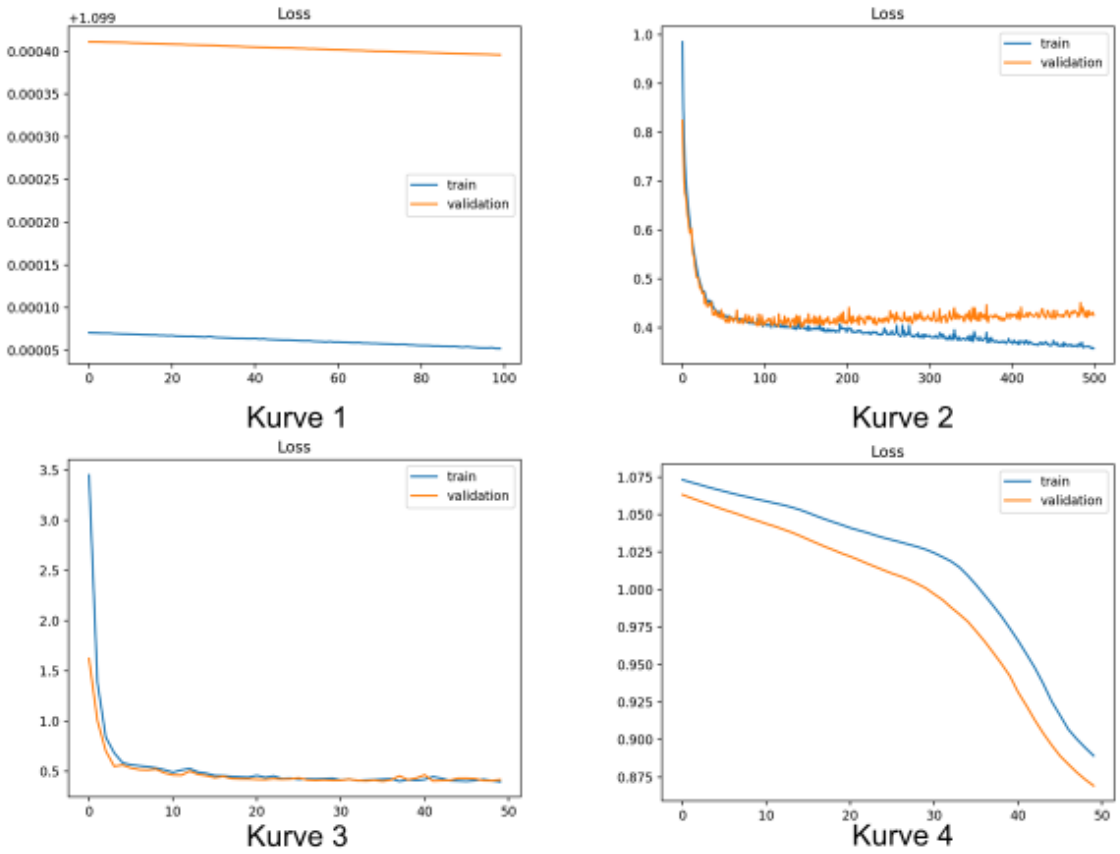

Abbildung 4, Beispiele für learning curves *mit* underfit *(Kurve 1),* overfit *(Kurve 2),* good fit *(Kurve 3) und unzureichende Trainingsdauer (Kurve 4) (https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/) (31)*

1.3.4 Anwendungsbereiche

Methoden des maschinellen Lernens kommen in vielen Bereichen des Alltags zum Einsatz. Die Funktion des CNN wird sich beispielsweise auch in der Spracherkennung zu Nutze gemacht, um sich so effektiven Übersetzungstools zu bedienen (34).

Im medizinischen Sektor kann maschinelles Lernen bei großen Mengen an medizinischen Bildgebungsdaten zum Tragen kommen. Im Rahmen von epidemiologischen Studien ist die MRT Bildgebung maßgeblich (35). Die erheblichen Datenmengen sind zeitaufwendig und damit kostspielig auszuwerten, wodurch

Verfahren, die einzelne Schritte wie beispielsweise Volumenbestimmung automatisieren, immer mehr in Fokus geraten. Die Implementierung einer effektiven Methode, um Organsegmentierung zu vollautomatisieren stellt einen der anfänglichen Schritte auf dem Weg zur maschinellen Bildanalyse und Läsionsdetektion dar (5, 20).

Dies ebnet den Weg für weitere Bildanalysen, durch die automatische Segmentierung ist es möglich gewaltige Datenmengen wirtschaftlich und effektiv zu bearbeiten. So können aus der Segmentierung beispielsweise die Größen der Organe ermittelt und auch bei Verlaufsuntersuchungen rasch verglichen werden. Aus den gewonnenen Informationen der Segmentierung können auch weitere Untersuchungen folgen, wie zum Beispiel die Bestimmung und Veränderung des Fettgehaltes.

Lavdas et al. konnte zeigen, dass sich ein CNN gegenüber anderen getesteten CF oder MA Konzepten durchgesetzt hat. In der Kohortenstudie KORA konnten bereits mit einem CNN Erfahrungen gesammelt werden (20), sodass die Wahl im Fall der NAKO auf ein CNN als Algorithmus für die automatische Organsegmentierung fiel.

2. Zielsetzung

Bevölkerungsstudien wie die NAKO, mit hoher Teilnehmerzahl, aufwendiger Bildgebung und damit verbunden großen Datenmengen, gelangen vermehrt in den Fokus. Diese Datenmengen auszuwerten und zu analysieren ist komplex und zeitaufwendig. Deshalb ist eine Automatisierung notwendig.

Die MRT-Untersuchungen sollen Grundlage für umfassende phänotypische Charakterisierung der Teilnehmer sein und Aufschluss über Prävalenzen von Vorstadien chronischer Erkrankungen geben und die Entwicklung neuer Methoden der Prävention unterstützen. Zudem stellen diese Daten eine Basis für weiterführende Studien zu verschiedensten epidemiologischen Fragestellungen dar (5).

Bestehende KI Methoden zeigen erhebliche Vorteile auf, Algorithmen wie CNN arbeiten mit höchster Präzision und können enorme Bilddatenmengen bewältigen. Nachteile sind unter anderem die manuelle Erstellung der Trainingsdaten und die Limitierung des Algorithmus durch Informationsverlust und Lernbegrenzung bei nicht ausreichendem Training. Das macht eine ausführliche Validierung der Trainingsprozedur unabdingbar.

Ziel dieser Arbeit war daher die Implementierung und anschließende Validierung der automatischen Organsegmentierung von Leber und Milz mittels CNN auf MRT Daten im Rahmen der NAKO Studie.

3. Methoden und Material

3.1 Studiendesign und zugrunde liegende Daten

Es liegt ein positives Ethikvotum zur Auswertung vor (Projekt-Nummer 120/2018BO2), sowie Aufklärungen und Einverständnisse aller Probanden, welche in den jeweiligen Studienzentren eingeholt wurden.

Bei den zur Verfügung gestellten Daten handelt es sich um MRT Ganzkörperaufnahmen, die im Rahmen der NAKO Gesundheitsstudie zwischen 2015 und 2016 akquiriert wurden. Diese liegen pseudonymisiert vor.

Retrospektiv wurden stichprobenartig 200 MRT Aufnahmen ausgewählt und bei insgesamt 100 Aufnahmen Leber und Milz manuell segmentiert. 80 als sogenannte Trainingsdatensätze, um die automatische Organsegmentierung mittels CNN zu etablieren und 20 als sogenannte Testdatensätze, die dann der Analyse der Genauigkeit des Algorithmus dienen (siehe Abbildung 5).

Zusätzlich lagen epidemiologische Daten zu Alter, Geschlecht und Body Mass Index (BMI) der Probanden vor.

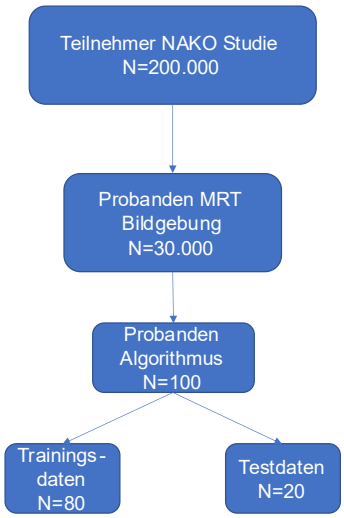

Abbildung 5, Schema zum Patientenkollektiv

3.2 Bildgebungsprotokoll

Bei den verwendeten MRT-Aufnahmen handelt es sich um T1 gewichtete *3D Volume Interpolated Breathhold Examination* (VIBE) 2-Punkt Dixon Sequenzen(36) aus 3.0-Tesla Ganzkörper-MRT Geräten (Magnetom Skyra von Siemens Healthineers) (5). Die Dauer der Gesamtuntersuchung betrug etwa eine Stunde, die Aufnahmedauer der hier analysierten Wichtung betrug etwa fünf Minuten. Es wurde kein Kontrastmittel verwendet (5).

Details des Bildgebungsprotokolls sind in Tabelle 1 aufgeführt.

Digital Imaging and Communications in Medicine (DICOM) Tags	
Beschreibung	Wert
Schichtdicke	3 mm
Wiederholungszeit	4,36 ms
Echozeit	1,36 ms
Pixelbandweite	975 Hz/px
Akquisitionsmatrix	320x208 Pixel
Flipwinkel	9 °
Zeilen	260
Spalten	320
Pixelabstand	1.40625 mm xw\1.40625 mm

Tabelle 1, DICOM Tags aus MicroDicom (DICOM Viewer) für Microsoft Windows(37)

3.3 Manuelle Organsegmentierung - Generierung von Trainingsdatensätzen

Zur Bearbeitung der Ganzkörper-MRT Aufnahmen wurde die Software *MITK Workbench 2015.5.2* (The Medical Imaging Interaction Toolkit, German Cancer Research Center, Division of Medical Image Computing, Heidelberg, Deutschland) für Microsoft Windows (38, 39) verwendet. Diese Software dient der Darstellung und Bearbeitung von medizinischen Bilddaten und ermöglicht außerdem nebst Volumenvisualisierung auch verschiedene Bild- und Messstatistiken (39).

3.3.1 Darstellung der Patientenaufnahmen im MITK

Der jeweils zu bearbeitende Datensatz wurde im MITK über den Pfad *DICOM -> Import -> Scan directory* importiert, ausgewählt und über *View* im Display dargestellt (siehe Abbildung 6).

Als Darstellung wurde die axiale Ansicht gewählt. Im Reiter *Segmentation* wurde im Feld *Data Selection* unter *Patient Image* der korrekt angezeigte Fall überprüft. Darunter im Unterfeld *Segmentation* über die Schaltfläche *Create a new segmentation* die zu führende Untersuchung mit *Liver* beziehungsweise *Spleen* tituliert und gespeichert.

Die Markierung der Organumrisse erfolgte mit der Funktion *Add* im Feld *2D Tools* für jedes einzelne Schnittbild. Korrekturen wurden mit der Funktion *Subtract*, im selben Feld, durchgeführt. Es wurde zunächst am oberen Leberpol begonnen und in jedem Schnittbild die Umrisse markiert. Die Vena Cava inferior (untere Hohlvene) wurde ausgespart. Gallenblase und Leberhilus wurden mitsegmentiert.

Für die Milz wurde auf gleiche Art und Weise vom unteren Milzpol kopfwärts bearbeitet. Es resultieren die manuellen Segmentierungen wie in Abbildung 7 dargestellt.

Abschließend sind die Segmentierungen von Leber und Milz in einem separaten Ordner auf einer verschlüsselten Festplatte gesichert worden.

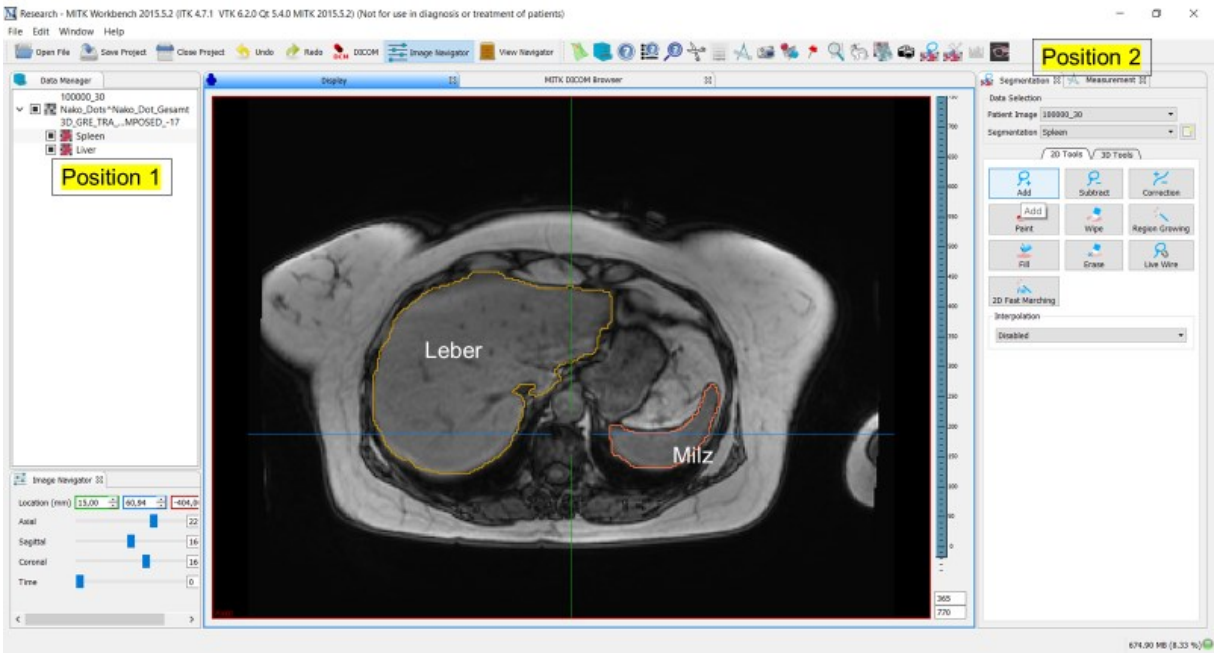

Abbildung 6, Beispielscreenshot mit segmentierten Leber- und Milzumrissen in der axialen Ebene (MITK für Microsoft Windows) Über Position 1 sind im Data Manager der Dateiname, in dem Fall die Patienten ID, sowie die Segmentationsdateien Spleen und Liver erkennbar.

Unter Position 2 sind der Reiter Segmentation, die Felder Patient Image und Segmentation (hier Spleen), sowie die Funktionen Add und Subtract, mit denen die Segmentierung durchgeführt wurde.

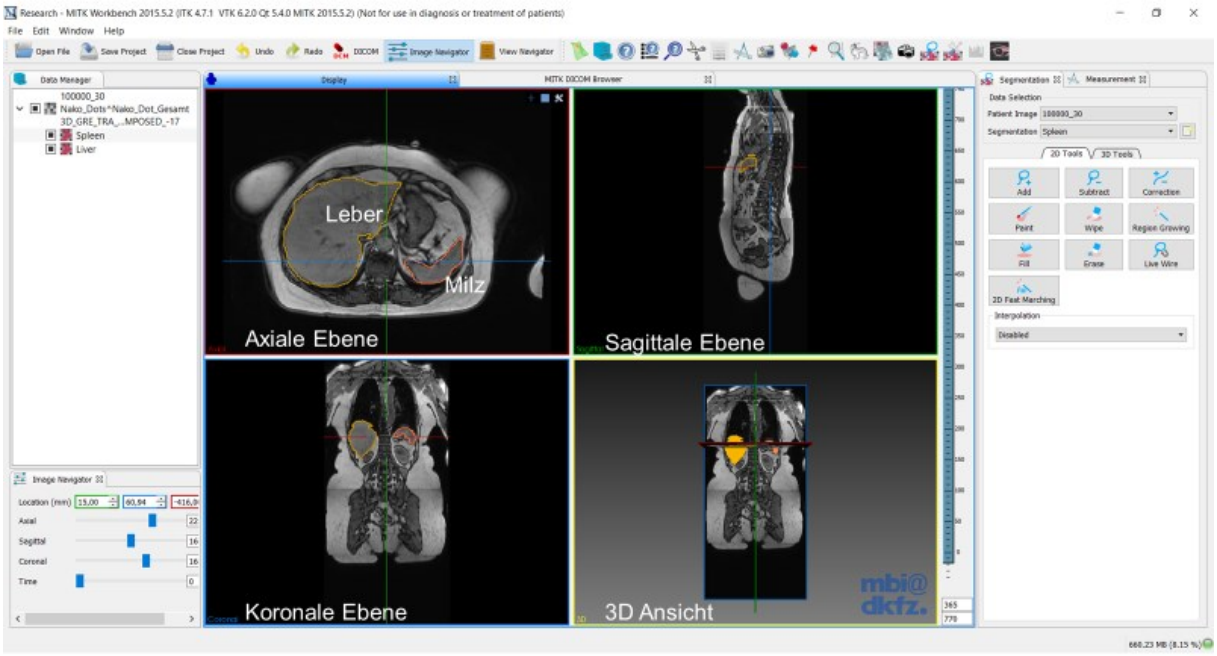

Abbildung 7, Beispielscreenshot mit verschiedenen Ebenen eines segmentierten Falles (MITK für Microsoft Windows)

Für die Feststellung der Dauer der manuellen Segmentierung wurde bei der Bearbeitung der Testdatensätze die Anfangs- und Endzeit sowohl der Leber – als auch der Milz Segmentierung festgehalten. Die Dauer der manuellen Segmentierung von Leber und Milz in MITK wurde mittels einfacher Stoppuhr gestoppt. Es wurde der Mittelwert der Standardabweichung (SD) der Dauer der manuellen Segmentierung für den jeweiligen Probanden bestimmt.

3.4 Automatische Organsegmentierung – Parameter des CNN Algorithmus

Die automatische Organsegmentierung durch ein *Deep Learning* Verfahren, den CNN Algorithmus, wurde anhand etablierter und validierter Grundkonzepte und Einstellungen vorgenommen (20). Der hier verwendete CNN Algorithmus besteht aus vielen verschiedenen, erprobten Konzepten. Diese Konzepte wurden mit Hilfe der Vorarbeit an der Kohortenstudie KORA ausgewählt.

Grafisch dargestellt erkennt man zwei Pfade, auf der Inputseite (links) steht der kontrahierende Teil, auf der Outputseite (rechts) der expandierende (siehe Abbildung 8). Jede Stufe besteht aus sogenannten N_B, für *dense blocks*, mit entsprechender

Layeranzahl L und N_T-Blöcken, für *Transition Layer Pool/Up*, untereinander verbunden mit *c* (einer vorwärts gerichteten horizontalen Verbindung zwischen den Pfaden) (20).

Verschiedene Konzepte und andere CNN wurden zu einem CNN kombiniert. Es gibt ein sogenanntes UNet Konzept, ein CNN Algorithmus, der speziell darauf ausgelegt ist, mit einer modifizierten, erweiterten Struktur eine geringere Anzahl an Trainingsdaten zu bearbeiten und dennoch genauere Segmentierungen zu erhalten (20, 25).

Zusätzlich wurden Aspekte eines VNet für Volumensegmentierung in medizinischen Bilddaten benutzt. Diese UNet und VNet sind geeignet für die Segmentierung und geben den ablaufenden sogenannten kontrahierenden und expandierenden Pfad eine U- bzw. V – förmige Form, wie in Abbildung 8 ersichtlich (20, 25, 26). Beide Pfade, der kontrahierende, sowie auch der expandierende bestehen aus mehreren *layers*. Im kontrahierenden Pfad geht auf Kosten von Kontext Ortsinformation verloren, als Ausgleich hierzu besteht eine vorwärts gerichtete horizontale Verbindung zum expandierenden Pfad. Dies ermöglicht im expandierenden Pfad eine kontextbasierte genaue Ortslokalisierung der zu bestimmenden Texturen (21).

Außerdem wurde die ResNet Architektur im UNET integriert, es ist ein Beispiel für ein DCNN und vergrößert das Netzwerk auf 152 *layers* (20, 40).

DenseNet soll zur Tiefenkontrolle dienen, dadurch werden alle *layers* untereinander verbunden, um Informationsverlust vorzubeugen (20, 41).

All diese Konzepte dienen als Grundlage für den CNN und bestimmen dessen Performance, durch Vorerfahrungen mit der Kohortenstudie KORA konnten die beschriebenen Konzepte genutzt werden, um den CNN Algorithmus zu optimieren und Fehler zu minimieren (20).

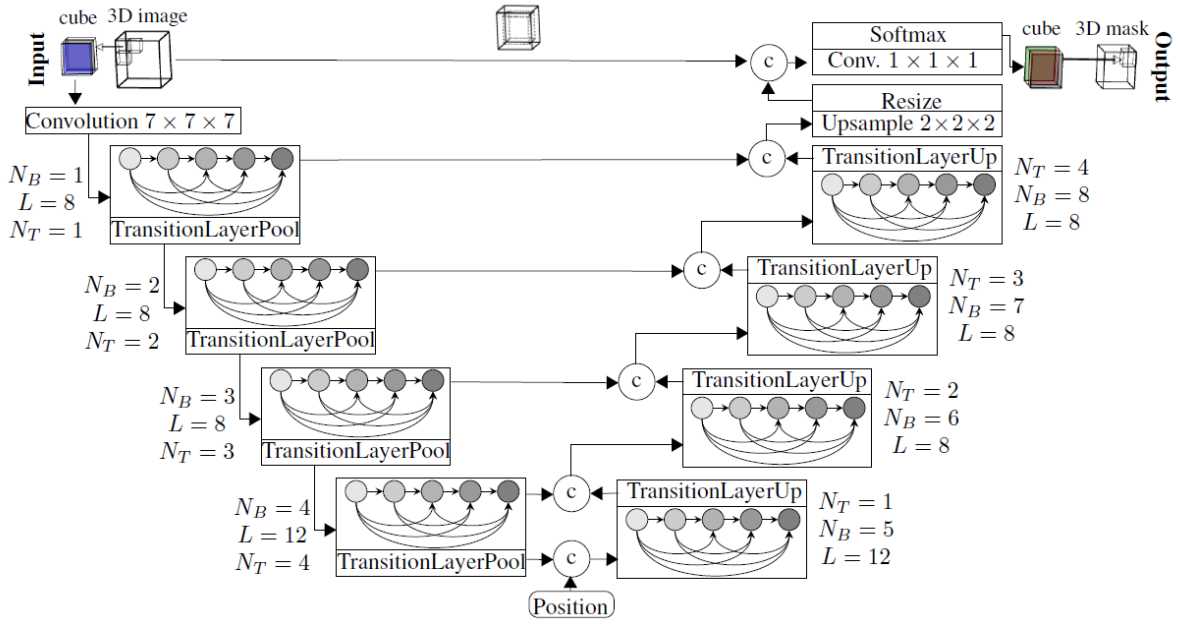

Abbildung 8, Darstellung des verwendeten CNN Ablaufs. Auf der Inputseite (links) steht der kontrahierende Teil, auf der Outputseite (rechts) der expandierende. Jede Stufe besteht aus sogenannten N_B, für dense blocks*, mit entsprechender Layeranzahl L und N_T-Blöcken, für* Transition Layer Pool/Up*, untereinander verbunden mit* c*(20).*

Für die Hardware des CNN Algorithmus wurde eine *GTX 1080 TI* Grafikkarte GPU verwendet. Als Hyperparameter wurden unter anderem eine *Buffer Capacity* von 15 Patienten, eine *Batch Size* von 48 und *Batches Per Shift* von 35 gewählt.

3.4.1 Trainingsprozedur

Das Training erfolgte auf 80 von 100 Datensätzen, welche zufällig ausgewählt wurden. Innerhalb dieses Trainingsdatensatzes erfolgte ein Hyperparametertuning mittels 5-facher *Cross-Validierung* zur Bestimmung der optimalen Anzahl an sogenannten *Epochen* (Anzahl an Trainingsdurchläufen). Bei jedem Trainingsschritt wurden somit 16 von 80 Datensätzen zur Validierung verwendet. Danach wurde mit allen 80 Datensätzen mit der optimalen Anzahl an *Epochen* trainiert.

Nach Abschluss des Trainingsvorganges wurden dem trainierten Algorithmus die verbleibenden 20 Datensätze dargeboten, also die vom Algorithmus noch ungesehenen Daten.

Zur Validierung des Lernprogresses des Algorithmus wurde die *loss curve* ermittelt und analysiert.

Die Dauer des Trainingsprozesses der automatischen Organsegmentierung, und auch die der Validierungsprozedur, konnte aus dem Ablauf des Algorithmus ausgelesen werden.

3.5 Statistische Auswertung

Die statistische Auswertung wurde mit *IBM SPSS Statistics 24* für Microsoft Windows durchgeführt. Zum statistischen Vergleich bezüglich Alters – und BMI – Verteilung im Patientenkollektiv wurde der t-test für unabhängige Stichproben verwendet, Die Untersuchung auf Normalverteilung erfolgte mit dem Kolmogorov-Smirnov-Test. Für Organvolumina und Mittelwerte der Fettgehälter wurden die Korrelationen nach Pearson bestimmt. Mittels t-test für unabhängige Stichproben wurde die Signifikanz für mittlere Fettgehalte und Organvolumina überprüft.

Zur Messung der Übereinstimmung wurden die Koeffizienten Dice und Jaccard verwendet. Des Weiteren wurden die Metriken Sensitivität, Spezifität, Präzision und die falsch negativ Rate ermittelt.

3.5.1 Untersuchung der Trainings- und Testdaten

Es wurden, unabhängig des *Deep Learning* Verfahrens, verschiedene erhobene epidemiologische Daten der 80 Trainings- und 20 Testprobanden verglichen. Die Daten für Probandenalter und BMI wurden nach Mittelwert und der Standardabweichung hin untersucht. Die Geschlechterverteilung wurde prozentual dargestellt. Darüber hinaus wurden die Mittelwerte und die Standardabweichung der manuell bestimmten Organvolumina von Leber und Milz verglichen, sowohl separat für die Trainings- und Testdaten, als auch für die gesamten 100 Probanden.

Es handelte sich nach dem Shapiro-Wilk-Test um normalverteilte Daten. Ein durchgeführter Levene-Test zeigte Varianzgleichheit. Demnach wurden die p-Werte mittels t-test für unabhängige Stichproben auf die Gleichheit der Mittelwerte untersucht. Bei der Geschlechterverteilung kam der Chi-Quadrat-Test zum Einsatz. Die graphische Darstellung der Organvolumina der zwei Organe erfolgte jeweils mittels Boxplots.

3.5.2 Untersuchung der manuellen und automatischen Organsegmentierung

Um die Trainingsprozedur zu evaluieren, wurden dem Algorithmus wie ebenda beschrieben 20 noch ungesehene Datensätze eingespeist. Es wurde untersucht, inwieweit das Organ erkannt und richtig markiert wurde. Die Resultate ließen sich mit den Ergebnissen der manuellen Organsegmentierung vergleichen. Hierzu wurde nach qualitativen, quantitativen und sekundären Aspekten untersucht.

3.5.2.1 Qualitative Bewertung der Organsegmentierung

Es wurden zwei verschiedene Klassifizierungsscores verwendet, um die Qualität der manuellen und automatischen Organsegmentierung qualitativ zu erfassen.

Zwecks qualitativer visueller Bewertung der automatischen Segmentierung von Leber und Milz in den Testdatensätzen wurde ein Score entworfen, um die grundsätzliche Segmentierungsleistung auf Vollständigkeit und Abweichungen hin visuell zu klassifizieren. Wie in Tabelle 2 ersichtlich wurden zur Klassifizierung vier Kategorien im Sinne einer Ordinalskala von *geringsten Abweichungen* bis *Organ nicht erkannt* definiert. Um als *geringste Abweichung* zu gelten, durften die visuellen Abweichungen nur geringfügig ausfallen. Wurden kleinerer Organteile ausgespart oder die Organgrenze geringfügig überschritten wurde die Auswertung als *geringe Abweichung* kategorisiert. Bei größeren Abweichungen, Grenzüberschreitungen zu Nachbarorganen oder dem Aussparen ganzer Organteile lag eine *wesentliche Abweichung vor.* Im schlechtesten Fall wurde ein *Organ nicht erkannt*.

Score für Organsegmentierung		
Bewertung	Ziffer	Kriterien
Geringste Abweichungen	1	Visuell geringfügig erfassbare Abweichungen
Geringe Abweichungen	2	Geringe Abweichungen, Aussparung kleinerer Teile, Organgrenzenüberschreitung
Wesentliche Abweichungen	3	Größere Abweichungen, Grenzverletzung benachbarter Organe, relevante Anteile fehlen, z.B. ein Leberlappen
Organ nicht erkannt	4	Organ nicht erkannt

Tabelle 2, Organscore für Vollständigkeit der Organsegmentierung

Um die Überschreitung der Organgrenze auf Segmentebene zu differenzieren wurde ein weiterer Score wie in Tabelle 3 eingeführt. Waren weniger als 10 Schichten

betroffen, lag eine *geringe*, bei mehr als 10 Schichten eine *größere* Fehlsegmentierung vor.

Score für Fehlsegmentierungen		
Bewertung	Ziffer	Kriterien
Gering	1	< 10 Schichten betroffen
Größer	2	> 10 Schichten betroffen

Tabelle 3, Score für Fehlsegmentierungen

Zur Veranschaulichung der Ergebnisse der beiden Scores wurden Balkendiagramme erstellt, welche die absoluten Häufigkeiten der Kategorien abbilden.

Um die automatisch segmentierten Daten mittels der erstellten Scores auswerten zu können, wurde das Programm *Aliza Medical Imaging & DICOM Viewer* (Bonn, Germany), für Microsoft Windows (42) benutzt. Das Ursprungs-MRT des jeweiligen Patienten wurde zusammen mit der automatisch segmentierten Organmaske in das Programm geladen. Dazu wurde der betroffene, unbearbeitete Datensatz zunächst in MITK geöffnet und im *.nii* Format (43) abgespeichert.

Im Aliza Viewer wurde dann die entsprechende MRT Ganzkörperaufnahme zusammen mit der automatisch erstellten Organmaske geöffnet (siehe Abbildung 9).

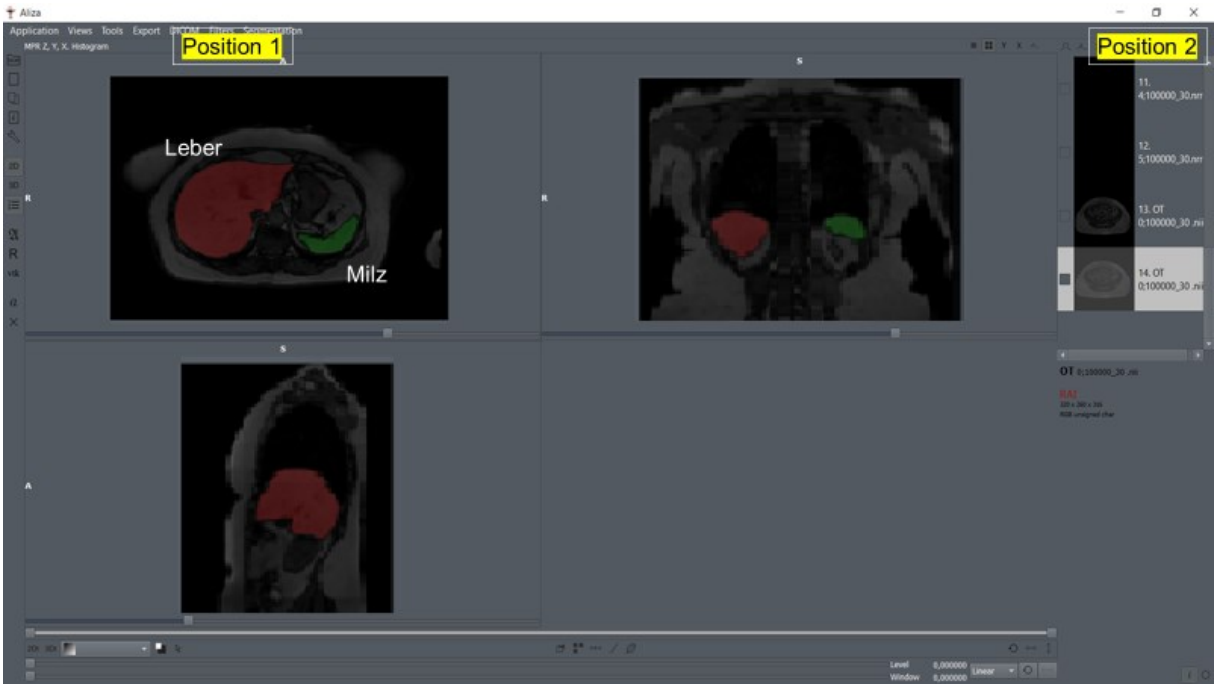

Abbildung 9, Beispielscreenshot mit hochgeladener MRT Aufnahme und segmentierter Organmaske (Aliza Viewer für Microsoft Windows). Bei Position 2 *wurde ein Bildmodus ausgewählt und mittels des Reiters* Filter *oberhalb von* Position 1 *über die Schaltfläche* Fusion/Extract physical space *mit der passenden Organmaske für jeweils Leber und Milz kombiniert und farblich dargestellt.*

3.5.2.2 Quantitative Analyse

Zur quantitativen Analyse der segmentierten Ergebnisse der 20 Testprobanden wurden verschiedene Metriken bestimmt. Die Sensitivität, die Spezifität, die Präzision, die Falsch-negativ-Rate und die Falsch-positiv Rate, sowie die Koeffizienten Dice und Jaccard. Dice und Jaccard sind statistische Werte für die Ähnlichkeitsanalyse und dienen dem Messen von Gleichheit und Diversität von Datensätzen (44).

Der Dice Koeffizienten wurde auf Minimum, Maximum, Mittelwert und Standardabweichung sowie Quartile als auch Median hin untersucht. Es handelte sich nach dem Shapiro-Wilk-Test um normalverteilte Daten. Ein durchgeführter Levene-Test zeigte Varianzgleichheit. Demnach wurden die p-Werte mittels t-test für unabhängige Stichproben auf die Gleichheit der Mittelwerte untersucht.

3.5.2.3 Vergleich sekundärer Bildinformationen

Aus den manuell und automatisch segmentierten MRT Aufnahmen wurden mehrere Daten erhoben. Die Segmentierung jedes aufeinander folgenden Schnittbildes erzielte eine dreidimensionale Markierung der Zielorgane. Dadurch ließen sich Organvolumina errechnen. Des Weiteren wurde der Fettgehalt der segmentierten Organe bestimmt. Hierzu bediente man sich Dixon-Sequenzen. Die Fettquantifizierung geht aus der Dixon Sequenz hervor, diese gibt vier Arten von Bildern her: *water only*, *fat only*, in-phase (Darstellung von Wasser und Fettgewebe) und *opposed-phase* Bilder (Darstellung von Wasser und Fettgewebe, Fettgewebe jedoch mit Signalabfall). Der Fettanteil wurde voxelweise innerhalb eines errechneten Bildes aus der *fat only* und *in-phase* Organmaske berechnet (36, 45). In Abbildung 10 und Abbildung 11 wurde ein Beispiel der Fettquantifizierung der manuellen und automatischen Organsegmentierung aufgezeigt.

Für die 20 Testprobanden wurden Organvolumina und mittlerer Fettgehalt, getrennt nach manueller und automatischer Segmentierung, für Leber und Milz untersucht und dargestellt.

Es handelte sich um nicht normalverteilte Daten, welche mit dem Wilcoxon-Test bei verbundenen Stichproben untersucht wurden. Tabellarisch wurden Minima, Maxima, Mittelwerte mit Standardabweichung sowie Median als auch Quartilenabstände für Organvolumen und mittleren Fettgehalt der Leber und Milz dargestellt. Für Leber und

Milz wurde die prozentuale Differenz als Mittelwert mit Standardabweichung für die jeweiligen Organvolumina und den mittleren Fettgehalt bestimmt. Es erfolgte eine graphische Veranschaulichung durch Boxplots, jeweils manuelle und automatische Segmentierung vergleichend. Zu Validierungszwecken wurden die manuell und automatisch ermittelten Organvolumina, sowie der mittlere Fettgehalt graphisch miteinander verglichen und auf Korrelation überprüft. Die Korrelation wurde nach Spearman-Rho ermittelt und die Signifikanz durch den p-Wert errechnet.

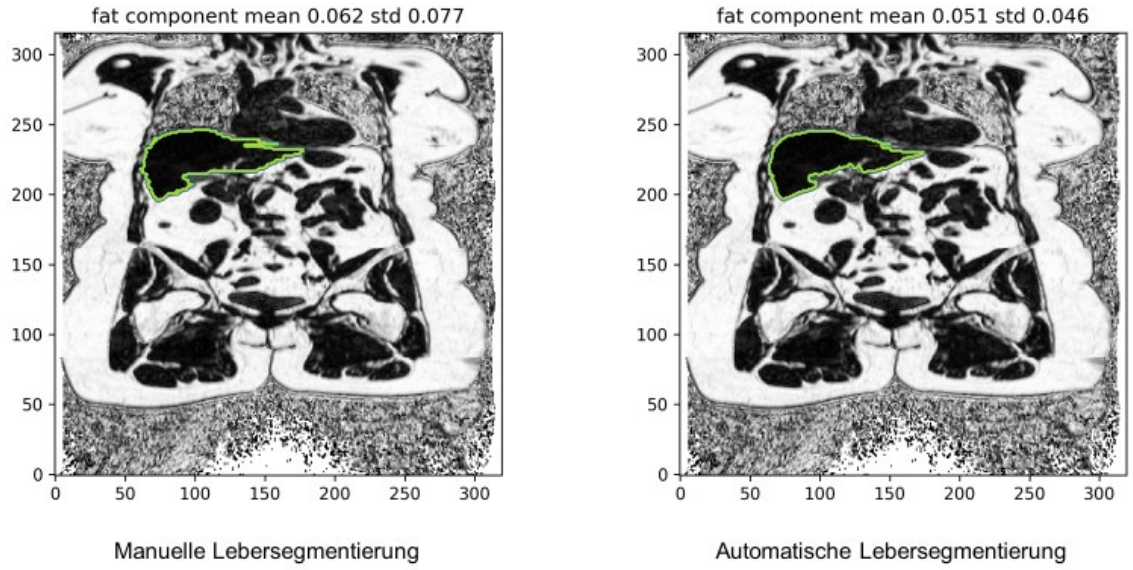

*Abbildung 10, Fettquantifizierung der Leber, Dixon Sequenz (*fat only*)*

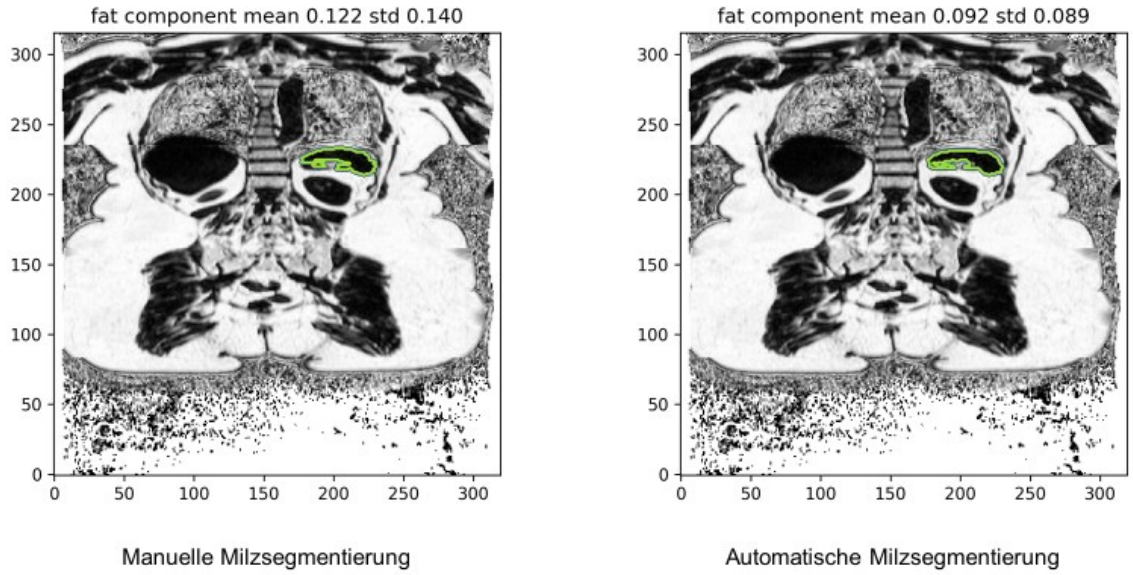

*Abbildung 11, Fettquantifizierung der Milz, Dixon Sequenz (*fat only*)*

4. Ergebnisse

4.1 Manuelle Organsegmentierung

Die Generierung der manuellen Segmentierungen war bei allen ausgewählten Probanden technisch umsetzbar. Die Organe Leber und Milz ließen sich problemlos segmentieren. Die Erstellung der Lebersegmentierung gestaltete sich zeitaufwendiger als die Segmentierung der Milz. Insbesondere die Erstreckung des Organs Leber auf wesentlich mehr Schichten im MRT, aber auch die Aussparung der unteren Hohlvene stellten sich als aufwendig und zeitintensiv dar.

Der Mittelwert der Segmentierungsdauer für die Leber betrug 36 Minuten. Die Standardabweichung betrug etwa 4 Minuten.

Für die Milz dauerte die Segmentierung im Mittelwert 6 Minuten mit einer Standardabweichung von rund 2 Minuten.

4.2 Automatische Organsegmentierung

Es zeigte sich, dass die automatische Organsegmentierung von Leber und Milz generell möglich war. Grundsätzlich wurden die Organe von dem angepassten CNN erkannt und ordnungsgemäß segmentiert.

4.2.1 Trainingsprozedur

Im ersten Trainingsschritt wurde die optimale Anzahl an benötigten *Epochen* bestimmt. Es wurden die 80 Datensätze mit jeweils 64 Trainingsdaten und 16 Validierungsdaten trainiert. Dabei wurden 50 *Epochen* benutzt, wobei eine *Epoche* etwa 35 Minuten veranschlagt hat. Somit betrug die erste Trainingsdauer (bei fünffacher *Cross-Validierung*) 5x50 *Epochen,* also circa fünf Tage.

In der in Abbildung 12 gezeigten *loss curve* zeigte sich eine optimale *Epochenzahl* von 40. Im anschließenden Training des endgültigen Modells mit allen 80 Datensätzen wurden 40 *Epochen* durchlaufen, was etwa 20 Stunden dauerte.

Das abschließende Modell wurde mit den 20 Testdaten evaluiert, pro Proband betrug die Messdauer zwischen fünf und zehn Minuten.

In diesen Zeiten wurden die Organe Leber, Milz und additiv Pankreas, linke und rechte Niere segmentiert (Pankreas und Nieren waren nicht Gegenstand dieser Arbeit).

Aus der Performance des Deep Learning Neural Network ließ sich während des Trainings mit den Trainingsdatensätzen die *loss curve* ablesen. Es zeigte sich ein typischer Verlauf für ein *good fit*, das bedeutet, dass Parameter und Trainingsdauer für den Algorithmus passend gewählt wurden. Sowohl im Training – als auch im Validierungsprozess wurde ein hoher Lernprogress verzeichnet, was sich durch das rasche Absinken der beiden Kurven, sowie durch die Annäherung an ein relativ niedriges Niveau auf der Y-Achse zeigt. Der Abstand der beiden Kurven zueinander sprach für eine hohe *generalization gap* (siehe Abbildung 12).

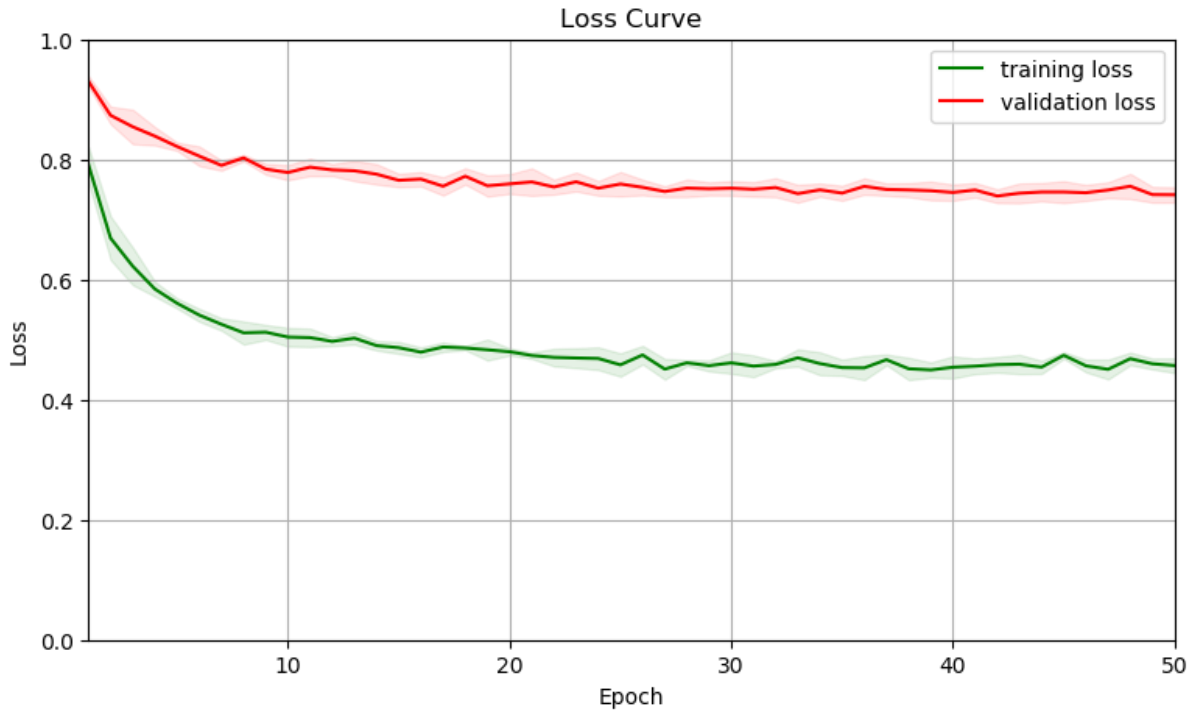

Abbildung 12, Beispiel einer loss curve *des verwendeten CNN Algorithmus.* training- *und* validation loss *senken sich auf ein bestimmtes Niveau ab und stabilisieren sich, der* training loss *bleibt unterhalb des* validation loss

4.3 Statistische Auswertung

4.3.1 Untersuchung der Trainings- und Testdaten

In Tabelle 4 sind das Trainings – und das Testkollektiv beschrieben.

Es zeigte sich in den untersuchten Trainings- und Testdaten die Ergebnisse bezüglich Geschlechts-, Alters- und BMI-Verteilung eine annähernde Gleichverteilung der Patientencharakteristika.

Es handelte sich um 38 männliche und 42 weibliche Studienteilnehmer. Das mittlere Alter lag bei 52 Jahren. Die Standardabweichung betrug 9,5 Jahre. Der BMI lag bei 27,0kg/m² mit einer Standardabweichung von 5,1kg/m².

Das Testkollektiv bestand aus den 20 manuell segmentierten MRT Aufnahmen von 11 männlichen und 9 weiblichen Probanden. Der Mittelwert des Alters der Gruppe lag bei 55 Jahren mit einer Standardabweichung von 9,8 Jahren. Der BMI lag bei 28,5kg/m², die Standardabweichung wurde mit 4,9kg/m² ermittelt.

Die manuell segmentierten Organvolumina von Leber und Milz zeigen im Trainings- und Testkollektiv ebenfalls nur geringe Abweichungen. Für die 100 Probanden zeigte sich ein Mittelwert des Organvolumens der Leber von 1,595l mit einer

Standardabweichung 0,368l, für die Milz wurde ein Mittelwert von 0,187l mit einer Standardabweichung von 0,086l errechnet. Dies ähnelt den Ergebnissen für das Trainings- respektive Testkollektiv.

In Tabelle 4 wurden die manuell segmentierten Organvolumina von Leber und Milz aufgeteilt in die Trainings- und Testdaten untersucht. In Abbildung 13 zeigten sich bei den Trainingsdaten einige einfache Ausreißer, in Abbildung 14 zeigte sich beim Trainingskollektiv ein einzelner einfacher Ausreißer, beim Testkollektiv ein extremer Ausreißer.

Patientenkollektiv				
		Datensatz N=100		p-Wert
		Training N=80	Test N=20	
		Mittelwert	Mittelwert	
Alter		52±9,5 Jahre	55±9,8 Jahre	.262
Geschlecht	Männlich	47,4%	55,5%	.548
	Weiblich	52,5%	45,0%	
BMI		27,0±5,1kg/m²	28,5±1,1kg/m²	.219
Organvolumen Leber		1,581±0,346l	1,653±0,451l	.717
Organvolumen Milz		0,181±0,075l	0,209±0,119l	.433

Tabelle 4, Kollektiv, bestehend aus Test – und Trainingsdatensätzen; Alter und Body-Mass-Index als Mittelwert und Geschlechteraufteilung in Prozent.

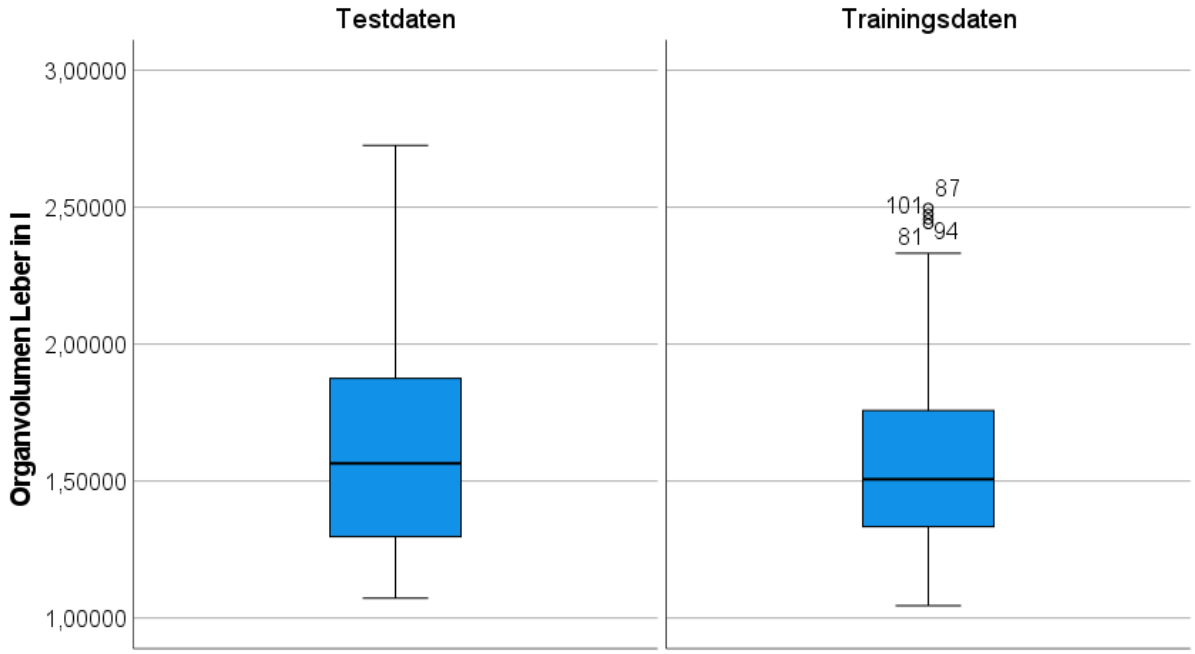

Abbildung 13, Organvolumina Leber für Trainings- und Testkollektiv

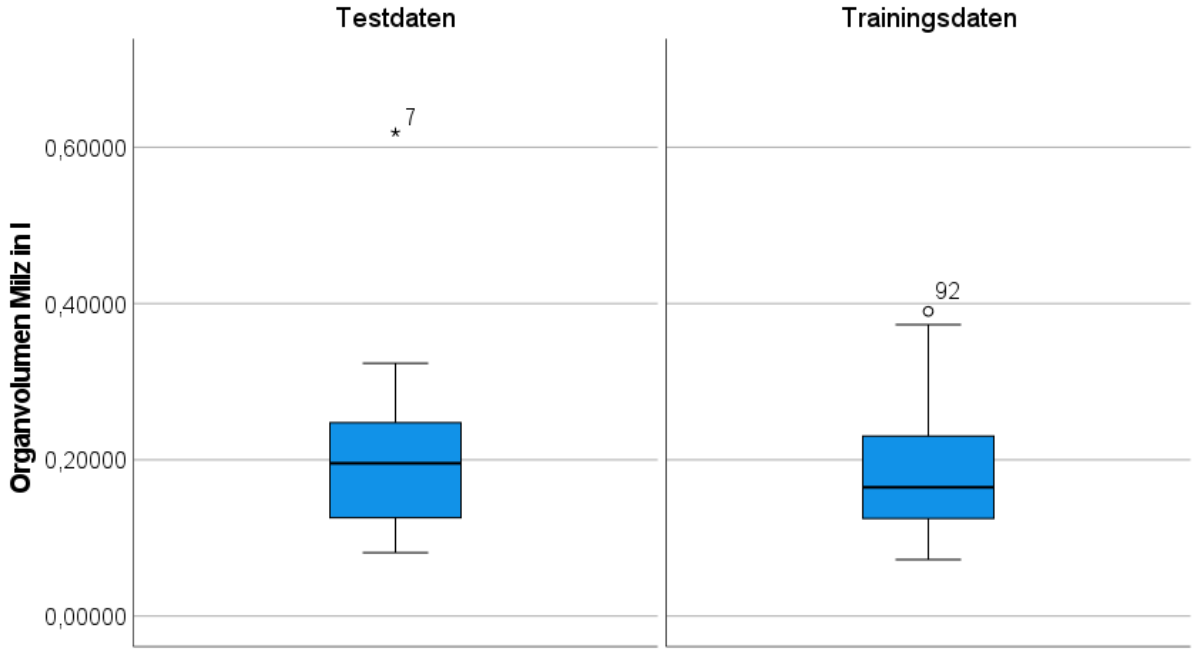

Abbildung 14, Organvolumina Milz für Trainings- und Testkollektiv

4.3.2 Untersuchung der manuellen und automatischen Organsegmentierung

Es zeigte sich, dass die Organe Leber und Milz grundsätzlich erkannt wurden. Es gab Unterschiede bei den Segmentierungen, vereinzelt wurden Gewebe segmentiert, die nicht zum Zielorgan gehörten, oder in einigen Fällen Organe unzureichend erfasst. Dies wird im Folgenden weiter untersucht.

4.3.2.1 Qualitative Bewertung der Organsegmentierung

Zur qualitativen Begutachtung der Segmentierungen wurden die Ergebnisse in der axialen und koronalen Ebene betrachtet. In den nachstehenden Abbildung 15 bis Abbildung 17 sind die Schnittbilder aller Testdaten dargestellt.

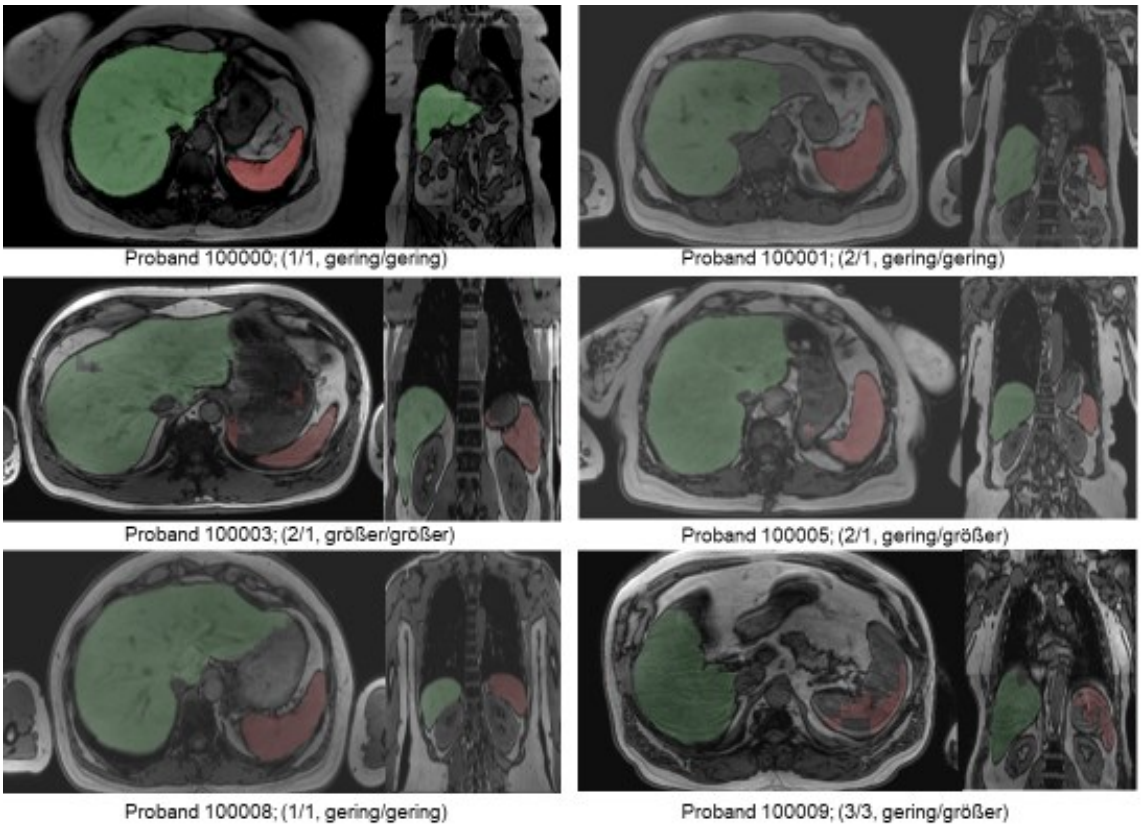

Abbildung 15, Probanden 100.000-100.009. Scoreergebnisse in Klammern nachgestellt: (Leber Score1/Milz Score1, Leber Score2/Milz Score2)

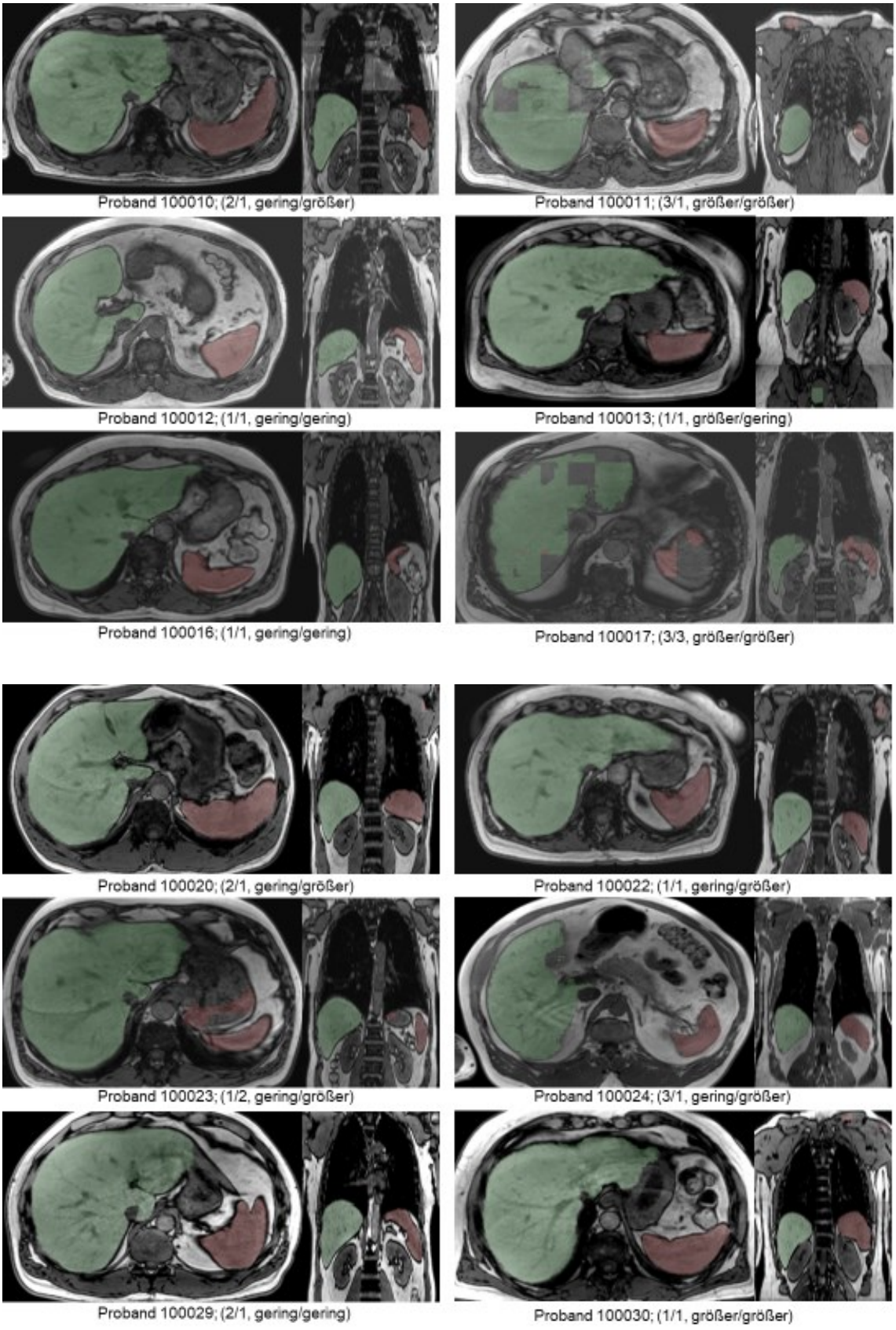

Abbildung 16, Probanden 100.010-100.030. Scoreergebnisse in Klammern nachgestellt: (Leber Score1/Milz Score1, Leber Score2/Milz Score2)

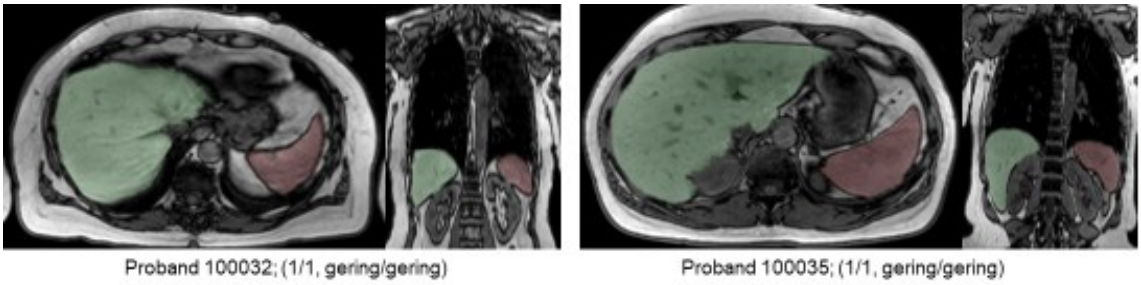

Abbildung 17, Probanden 100.030-100.035. Scoreergebnisse in Klammern nachgestellt: (Leber Score1/Milz Score1, Leber Score2/Milz Score2)

Es sind sowohl Beispiele für hochplatzierte Scoreergebnisse wie bei Proband 100.000, als auch Resultate mit niedrigplatzierten Scoreeinteilungen wie bei Proband 100.017 dargestellt.

Es konnte gezeigt werden, dass die Leber grundsätzlich erkannt wurde. In der automatischen Segmentierung wurde wie in den manuell segmentierten, eingespeisten Daten auch die untere Hohlvene ausgespart und Hilus und Gallenblase segmentiert. Zehn Segmentierungen erreichen die höchste Kategorie mit *geringste Abweichungen*. Es traten in der Kategorie *geringe Abweichungen* sechs Fälle auf, in der Kategorie *wesentliche Abweichungen* vier Fälle (siehe Abbildung 18).

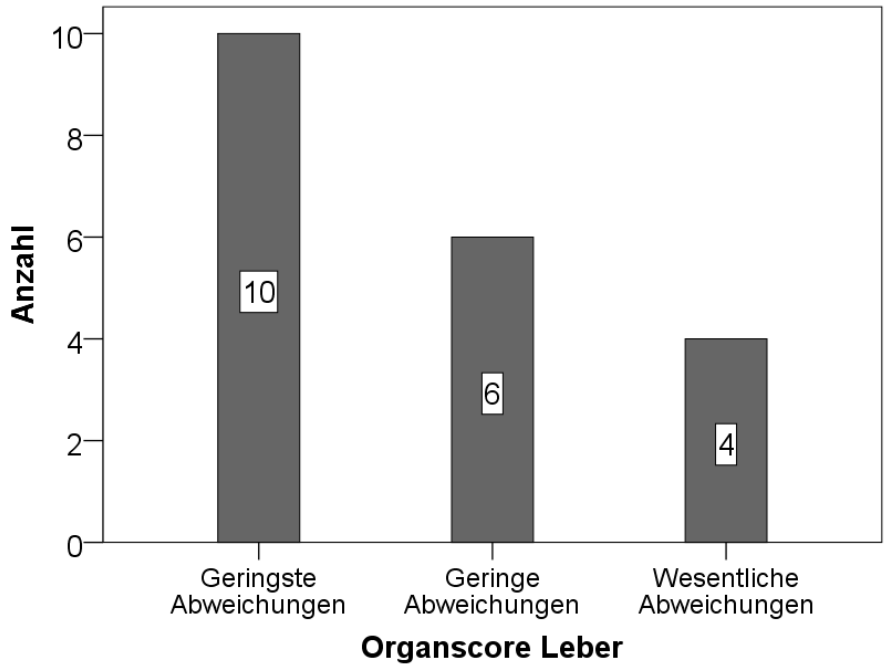

Abbildung 18, Kategorisierung der Ergebnisse für Leber in Organscore

Bei der qualitativen Auswertung der automatischen Milzsegmentierung wurde in jedem Fall das Organ erkannt. Es konnten 17 Fälle der Kategorie *geringste Abweichungen* zugeordnet werden. Ein bzw. zwei Fälle wurden *geringe Abweichungen* bzw. *wesentliche Abweichungen* zugeordnet (siehe Abbildung 19).

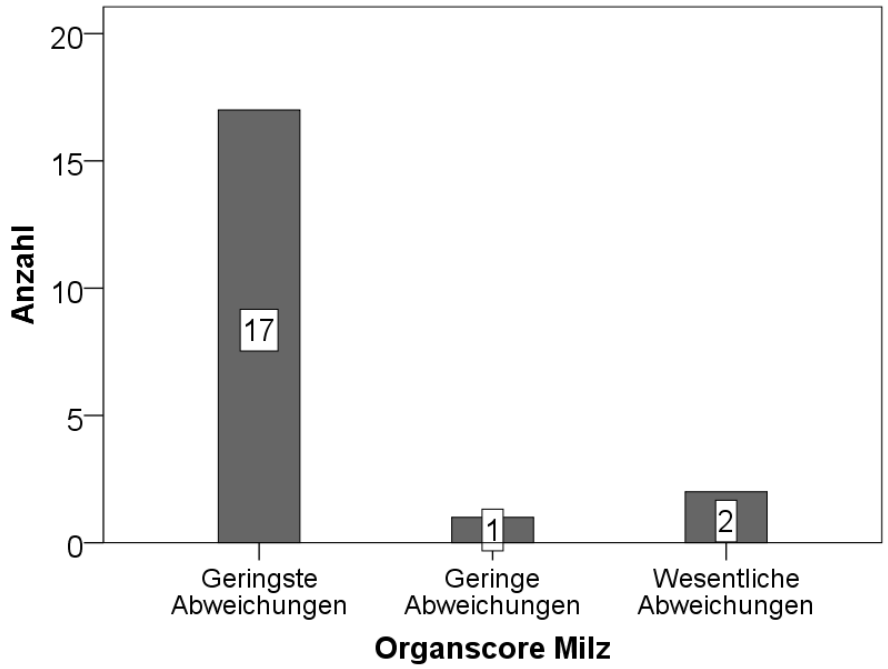

Abbildung 19, Kategorisierung der Ergebnisse für Milz in Organscore

In den gesichteten automatischen Lebersegmentierungen ließen sich 14 Segmentierungen in die Kategorie *gering* einstufen, das bedeutet sie wiesen in weniger als zehn Schnittbildern Segmentierungen außerhalb der Organgrenzen auf. Bei sechs Segmentierungen waren mehr als zehn Schnittbilder betroffen (siehe Abbildung 20).

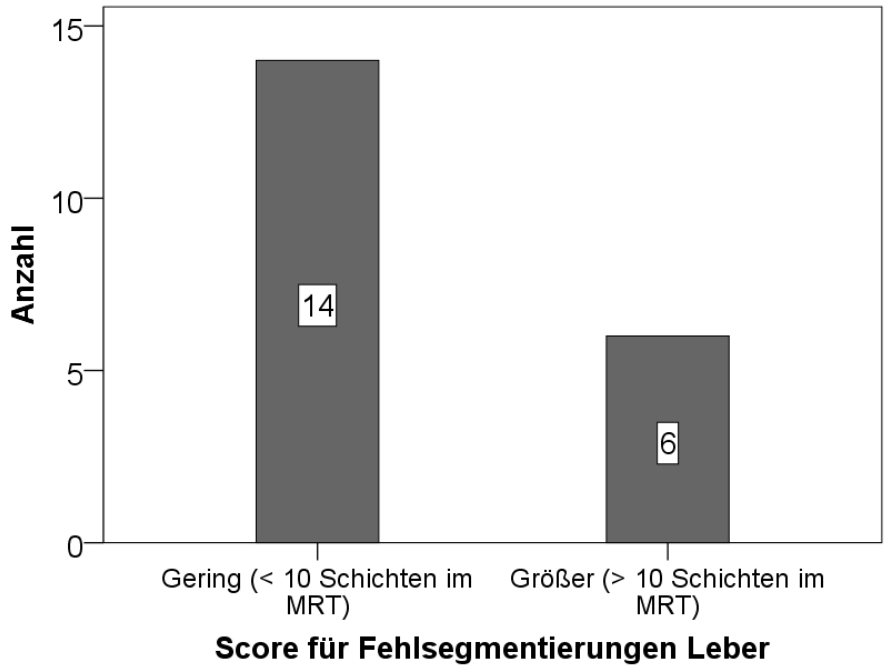

Abbildung 20, Kategorisierung der Ergebnisse der Leber, Score für Fehlsegmentierungen

Beim Organ Milz zeigte sich eine Gleichverteilung der Probanden (siehe Abbildung 21).

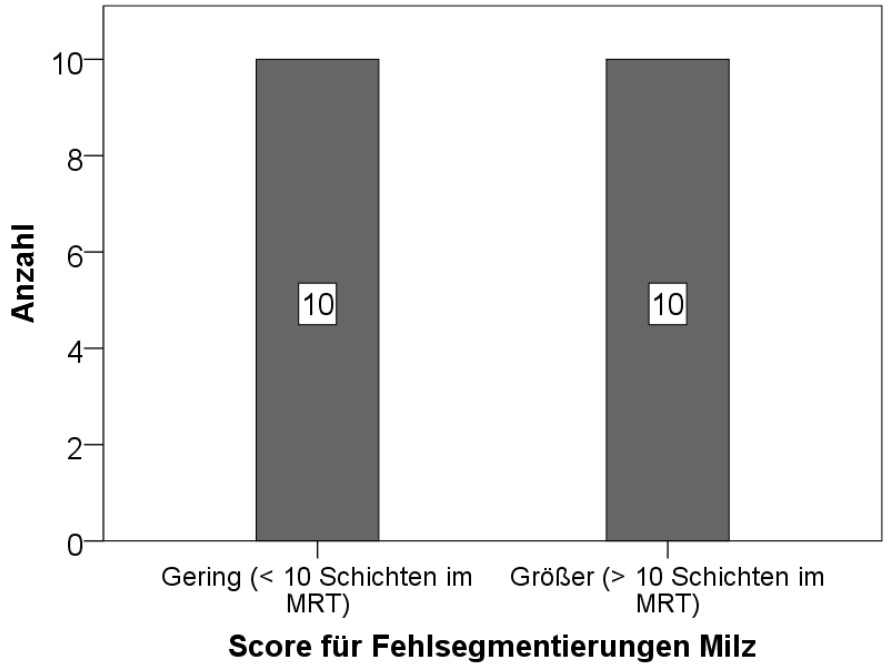

Abbildung 21, Kategorisierung der Ergebnisse für Milz in Score für Fehlsegmentierungen

Ein als gut bewertetes Beispielergebnis mit jeweiligen Bestqualifizierungen in beiden Scores zeigt sich in nachstehender Abbildung 22. Abbildung 23 zeigt ein Beispiel, in dem die Milz durch den Algorithmus nur mäßig gut segmentiert worden ist, im Organscore wurde das Ergebnis *wesentliche Abweichungen* und im Score für Fehlsegmentierungen das Ergebnis *größer* für mehr als zehn Schichten im MRT, bei denen fälschlicherweise organfremdes Gewebe segmentiert wurde.

Es zeigte sich, dass bei der Lebersegmentierung oftmals würfelartige Aussparungen innerhalb der Organgrenzen auftrafen. Eine Verfehlung der Organgrenzen trat jedoch weniger häufig als beim Organ Milz auf. Das Organ Milz wies ebenfalls die ähnlichen Aussparungen wie bei der Leber auf, jedoch wurde die Milz qualitativ besser segmentiert (Abbildung 18 und Abbildung 19). Es fiel auf, dass der Algorithmus bei der Milzsegmentierung öfter Teile fremder Gewebe miterfasste. So zeigte sich im Rahmen der qualitativen Analyse vermehrt segmentiertes Gewebe im Bereich der Schultermuskulatur und des Magens.

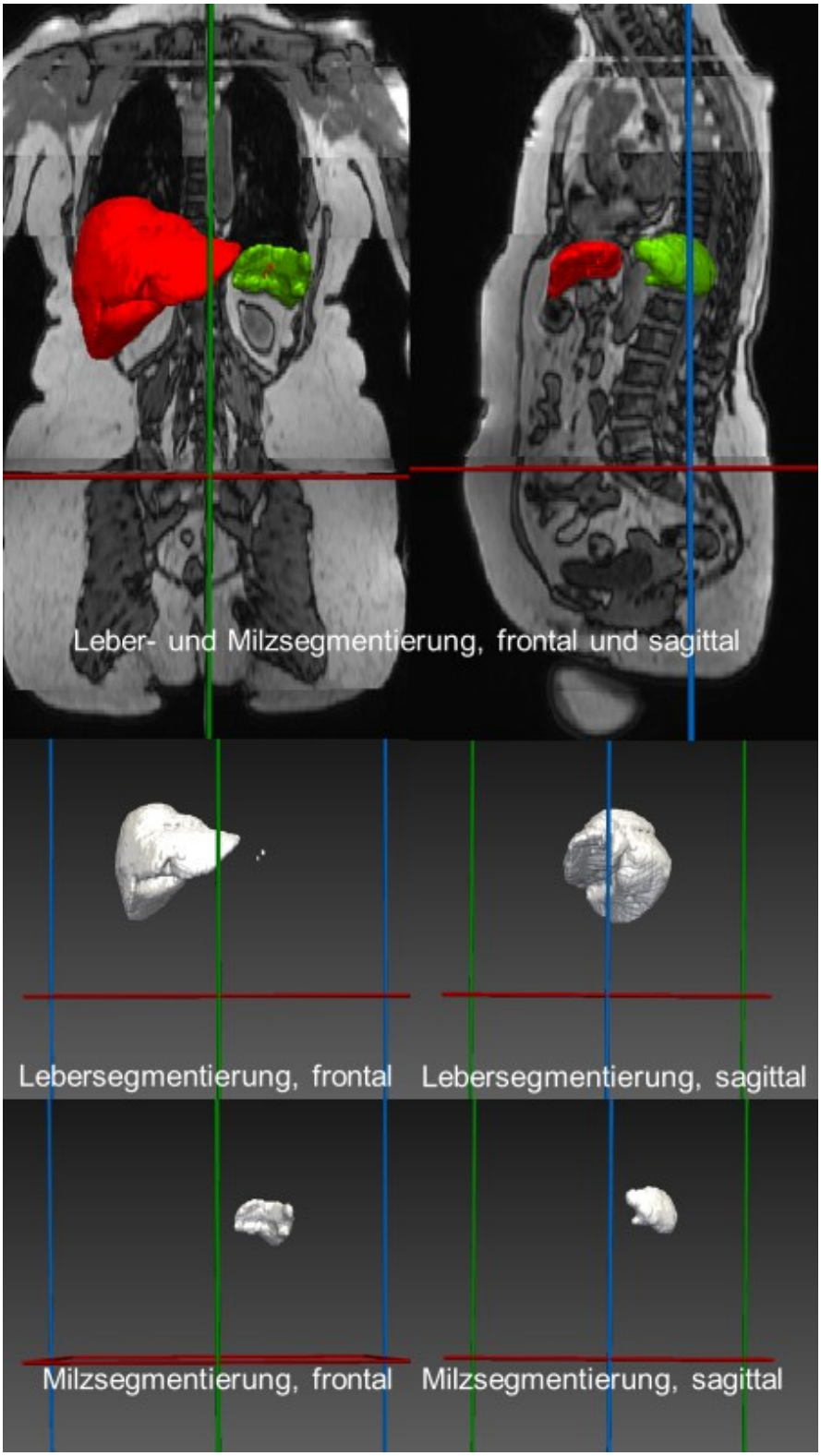

Abbildung 22, Beispielscreenshot mit hochgeladener MRT Aufnahme und 3D Darstellung der automatischen Organsegmentierung (MITK für Microsoft Windows). Es zeigt sich in dem Probanden ein qualitativ gutplatziertes Segmentierungsergebnis. Leber und Milz erreichen jeweils die höchste Kategorisierung in beiden Scores.

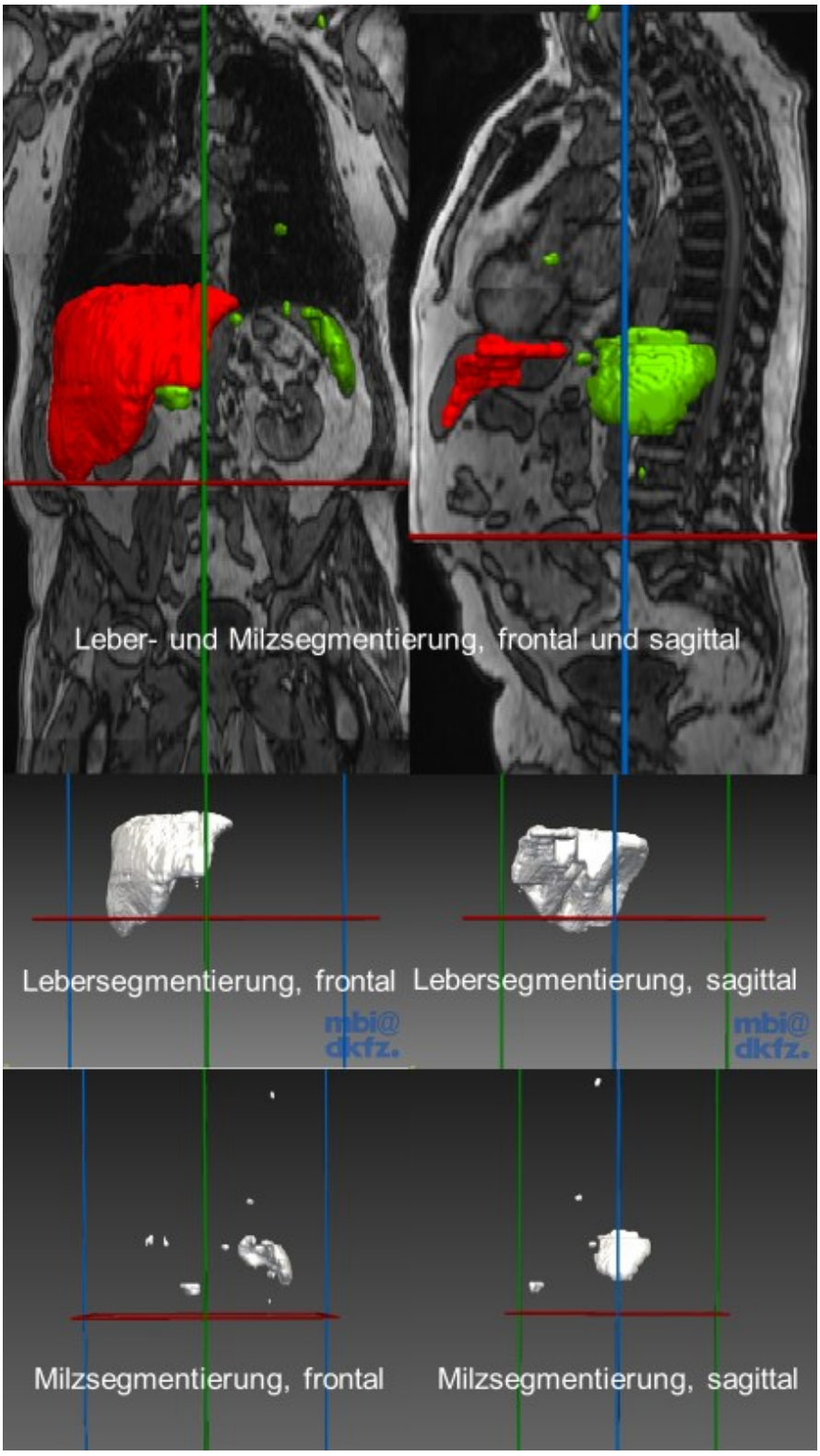

Abbildung 23, Beispielscreenshot mit hochgeladener MRT Aufnahme und 3D Darstellung der automatischen Organsegmentierung (MITK für Microsoft Windows). In den Kategorien Organscore Wesentliche Abweichungen *und* Größer *beim Score für Fehlsegmentierungen für das automatisch segmentierte Organ Milz.*

4.3.2.2 Quantitative Analyse

Bei der automatischen Organsegmentierung konnten hohe Übereinstimmungen bei sowohl Leber als auch Milzsegmentierungen festgestellt werden. In Tabelle 5 sind die Mittelwerte der Parameter zur Evaluation der Quantität des CNN aufgeführt. Es zeigten sich bis auf wenige Ausnahmen durchweg hohe Übereinstimmungen der ermittelten Werte.

Die Sensitivität des Algorithmus liegt für das Organ Leber bei 93,7%±3,5% und damit oberhalb des Wertes für das Organ Milz bei 84,8%±11,3%. Einen grundsätzlichen Unterschied der beiden Organsegmentierungen trat beispielsweise auch bei der Präzision des Algorithmus auf. Die Lebersegmentierung erreichte Werte von 94,1%±2,1% und die Milzsegmentierung niedrigere Werte mit 65,9%±15,7%. Insgesamt liegt die Präzision des verwendeten CNN Algorithmus bei 0,9960.

Die Leber zeigte eine falsch negativ Rate von 6%, die Milz von 15%. Damit wurde die Milz öfter außerhalb der Organgrenzen segmentiert.

Dies spiegelte sich bei den Koeffizienten Dice und Jaccard ebenfalls wider. In Tabelle 6 sind Minimum, Maximum und Mittelwert mit Standardabweichung des Dice Koeffizienten dargestellt. Außerdem wurden die Quartilenabstände und der Median bestimmt.

Bei der Lebersegmentierung ergaben sich ein Dice Koeffizient von 0,9393±0,0188 und ein Jaccard Wert von 0,8856±0,0326. Der Dice Koeffizient zeigt ein Minimum von 0,88 und ein Maximum von 0,96.

Die Koeffizienten Dice und Jaccard wurden bei der Milzsegmentierung mit jeweils 0,7423±0,1231 und 0,5902±0,1552 errechnet. Der Dice Koeffizient zeigt ein Minimum von 0,544 und ein Maximum von 0,923.

Für das Organ Leber wurde somit eine hohe Übereinstimmung der automatischen mit der manuellen Segmentierung durch die beiden angegebenen Koeffizienten errechnet. Der Algorithmus erreicht bei der Milz ebenfalls sehr hohe Werte, der Dice Koeffizient verdeutlicht aber einen Unterschied zwischen beiden Organsegmentierungen.

Evaluationsparameter des CNN				
Metrik	Hintergrund	Leber	Milz	Gesamt
Sensitivität	99,79±0,11%	93,77±3,50%	84,85±11,34%	
Spezifität	90,82±4,53%	99,91±0,04%	99,91±0,08%	
Präzision	99,80±0,16%	94,10±2,10%	65,97±15,72%	
Dice	0,9979±0,0011	0,9393±0,0188	0,7423±0,1231	
Jaccard	0,9959±0,0022	0,8856±0,0326	0,5902±0,1552	
Akkuratheit				0,9960

Tabelle 5, Mittelwerte±SD der Evaluationsparameter des CNN

Metrik	Lebersegmentierung	Milzsegmentierung	p-Wert
Minimum	0,88	0,54	
Maximum	0,96	0,92	
Mittelwert±SD	0,94±0,01	0,77±0,12	.457
Quartil 0,25	0,93	0,71	
Median	0,95	0,81	
Quartil 0,75	0,95	0,87	

Tabelle 6, Minimum, Maximum, Mittelwert±SD, Median und Quartile des Dice Koeffizienten der automatischen Organsegmentierung von Leber und Milz der 20 Testprobanden

4.3.2.3 Vergleich sekundärer Bildinformationen

Zunächst werden die Ergebnisse für das Organ Leber, dann das Organ Milz dargestellt.

In Tabelle 7 sind für die 20 Testprobanden das Minimum, Maximum, der Mittelwert mit Standardabweichung sowie Quartilen als auch Median des Organvolumens für die Leber aufgeführt, jeweils für die manuelle und automatische Organsegmentierung. In den Sekundärparametern wurde deutlich, dass zwischen manueller und automatischer Segmentierung nur geringfügige Abweichungen bestehen. Die Mittelwerte der Organvolumina waren mit 1,65l±0,45l und 1,58l±0,38l ähnlich. Es gab geringe Abweichungen durch ein leicht erhöhtes Maximum bei der automatischen Segmentierung mit 2,3l im Vergleich zur manuellen Segmentierung mit 2,72l. Der Median war bei der manuellen und automatischen Lebersegmentierung mit 1,57l gleich. Es zeigten sich Unterschiede in den Quartilen. Das untere Quartil ergab für die manuelle Segmentierung 1,28l, für die automatische 1,24l. Das obere Quartil lag bei 1,88l und bei 1,84l.

Die prozentuale Differenz zwischen dem manuell und automatisch segmentierten Organvolumen der Leber betrug 7,12%±12,26% (p-Wert .009).

Metrik	Volumen Leber Manuelle Segmentierung	Volumen Leber Automatische Segmentierung	p-Wert
Minimum	1,07l	1,01l	
Maximum	2,72l	2,3l	
Mittelwert±SD	1,65±0,45l	1,58±0,38l	.023
Quartil 0,25	1,28l	1,24l	
Median	1,57l	1,57l	
Quartil 0,75	1,88l	1,84l	

Tabelle 7, Vergleich Organvolumina Leber

In Tabelle 8 sind Minimum, Maximum, Mittelwert samt Standardabweichung sowie Quartile und Median für das resultierende Organvolumen bei der Milzsegmentierung aufgeführt. Auch hier zeigen sich ähnliche Mittelwerte mit 0,2l±0,11l bei der manuellen Segmentierung und 0,21l±0,09l bei der automatischen Organsegmentierung. Das Maximum von 0,61l bei der manuellen Segmentierung ist im Vergleich zum Maximum der automatischen Milzsegmentierung von 0,35l erhöht. Die untere Quartile zeigte 0,12l für die manuelle Segmentierung und 0,14l für die automatische. Der Median lag bei 0,2l, beziehungsweise 0,23l. Das obere Quartil wurde mit 0,25l und 0,28l errechnet.

Die prozentuale Differenz zwischen dem manuell und automatisch segmentierten Organvolumen der Milz betrug 0,72%±7,26% (p-Wert .331).

Metrik	Volumen Milz Manuelle Segmentierung	Volumen Milz Automatische Segmentierung	p-Wert
Minimum	0,08l	0,07l	
Maximum	0,61l	0,35l	
Mittelwert±SD	0,2±0,11l	0,21±0,09l	.05
Quartil 0,25	0,12l	0,14l	
Median	0,2l	0,23l	
Quartil 0,75	0,25l	0,28l	

Tabelle 8, Vergleich Organvolumina Milz

Exemplarisch sind die manuell und automatisch bestimmten Lebervolumina in Abbildung 24 dargestellt. Es zeigten sich die unterschiedlichen Maxima. Es wurden keine statistischen Ausreißer erfasst. Abbildung 25 wurde das in der manuellen und automatischen Segmentierung errechnete Organvolumen für die Milz verglichen. Hier zeigt sich wie in Tabelle 8 ersichtlich ein statistischer Ausreißer in der manuellen Segmentierung.

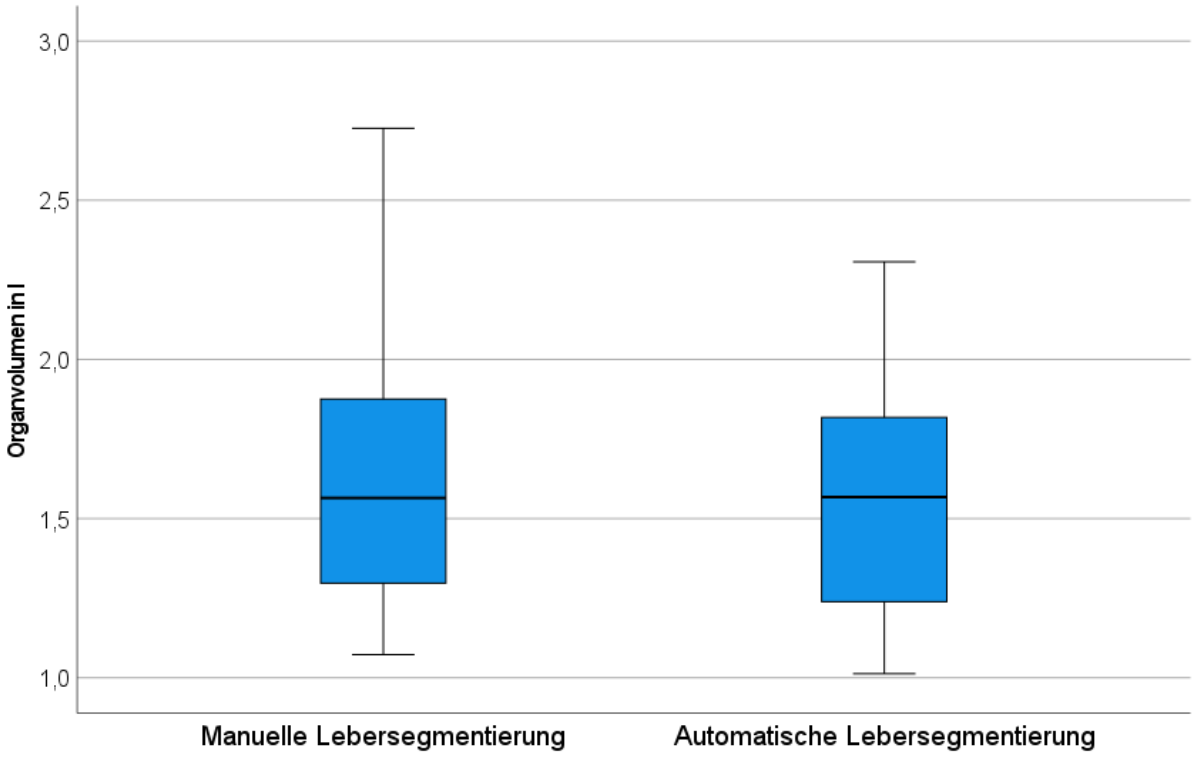

Abbildung 24, Vergleich Organvolumina Leber

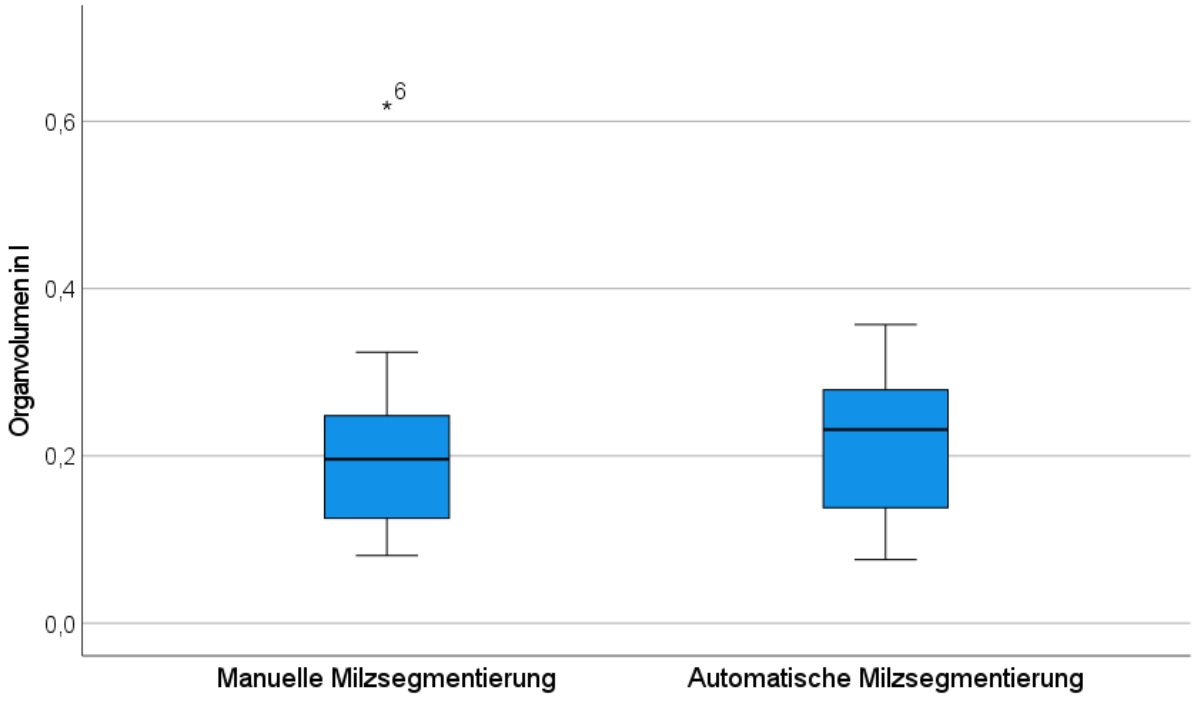

Abbildung 25, Vergleich Organvolumina Milz

In Tabelle 9 wurden Minimum, Maximum, Mittelwert mit Standardabweichung sowie Quartile und Median des mittleren Fettgehaltes der Leber dargestellt. Es erfolgte eine Unterscheidung nach manueller und automatischer Organsegmentierung. Es zeigte sich, wie auch schon beim Vergleich des Organvolumens, dass sich nur geringfügige Unterschiede im Fettgehalt widerspiegeln.

Die prozentuale Differenz zwischen dem mittleren Fettgehalt aus der manuellen und automatischen Segmentierung der Leber betrug 0,54%±0,34% (p-Wert <.001).

Metrik	Fettgehalt Leber Manuelle Segmentierung	Fettgehalt Leber Automatische Segmentierung	p-Wert
Minimum	4,3%	4,0%	
Maximum	24,8%	24,0%	
Mittelwert±SD	8,87±6,07%	8,33±6,05%	<.001
Quartil 0,25	5,1%	4,7%	
Median	6,3%	5,6%	
Quartil 0,75	10,1%	9,2%	

Tabelle 9, Vergleich Fettgehalt Leber

Tabelle 10 zeigt Minimum, Maximum, Mittelwert mit Standardabweichung sowie Quartile als auch Median für den mittleren Fettgehalt für die Segmentierungen des Organ Milz. Es zeigte sich ein erhöhtes Maximum in der manuellen Segmentierung mit 19,6% im Vergleich zur automatischen Segmentierung mit 16,0%. Ansonsten wurden nur geringe Unterschiede deutlich.

Die prozentuale Differenz zwischen dem mittleren Fettgehalt aus der manuellen und automatischen Segmentierung der Milz betrug 0,68%±1,31% (p-Wert .016).

Metrik	Fettgehalt Milz Manuelle Segmentierung	Fettgehalt Milz Automatische Segmentierung	p-Wert
Minimum	7,7%	7,1%	
Maximum	19,6%	16,0%	
Mittelwert±SD	11,38±3,26%	10,7±2,49%	.046
Quartil 0,25	8,9%	9,2%	
Median	10,7%	9,9%	
Quartil 0,75	13,1%	11,7%	

Tabelle 10, Vergleich Fettgehalt Milz

Exemplarisch wurde der mittlere Fettgehalt der manuellen und automatischen Organsegmentierung für die Organe Leber und Milz in Abbildung 26 und Abbildung 27. In Abbildung 26 zeigten sich statistische Ausreißer in ähnlichem Ausmaß in beiden Hälften des Diagramms, diese projizierten sich auf annähernd gleicher Höhe, was ebenfalls für hohe Übereinstimmungen spricht.

In Abbildung 27 werden Unterschiede deutlich. Die erste Diagrammhälfte zeigt insgesamt eine höhere Spannweite als das nebenstehende Diagramm, sowie einen statistischen Ausreißer. Das Diagramm der automatischen Milzsegmentierung weist ebenfalls einige statistische Ausreißer vor.

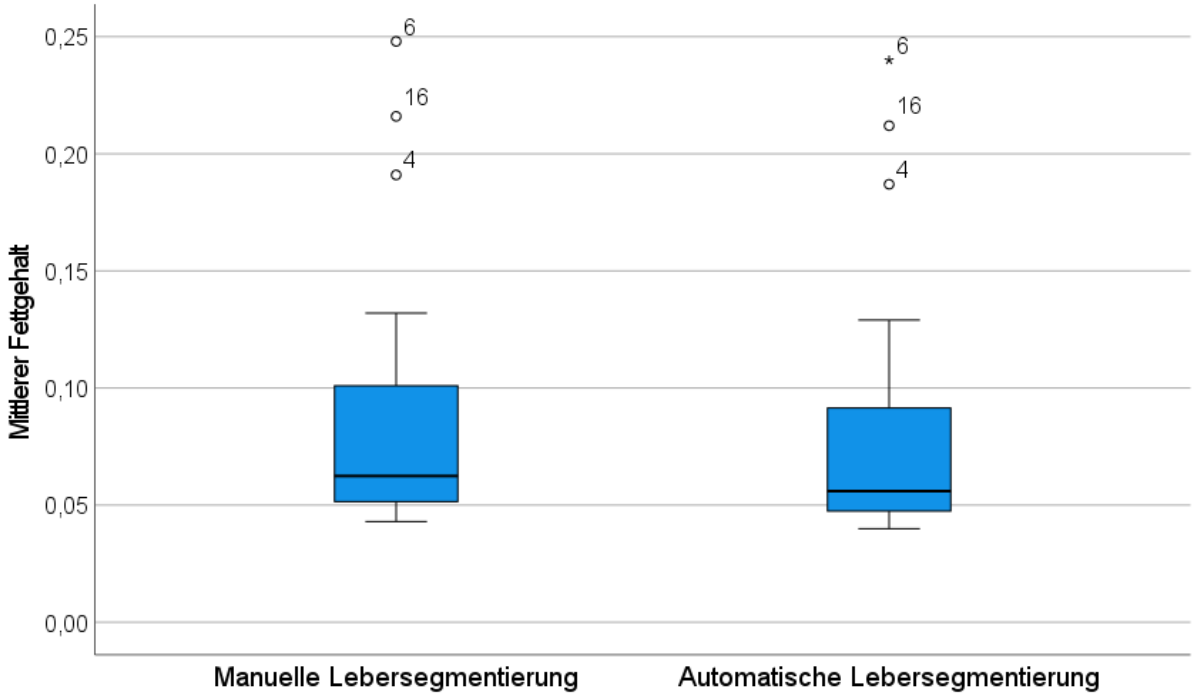

Abbildung 26, Vergleich Fettgehalt Leber

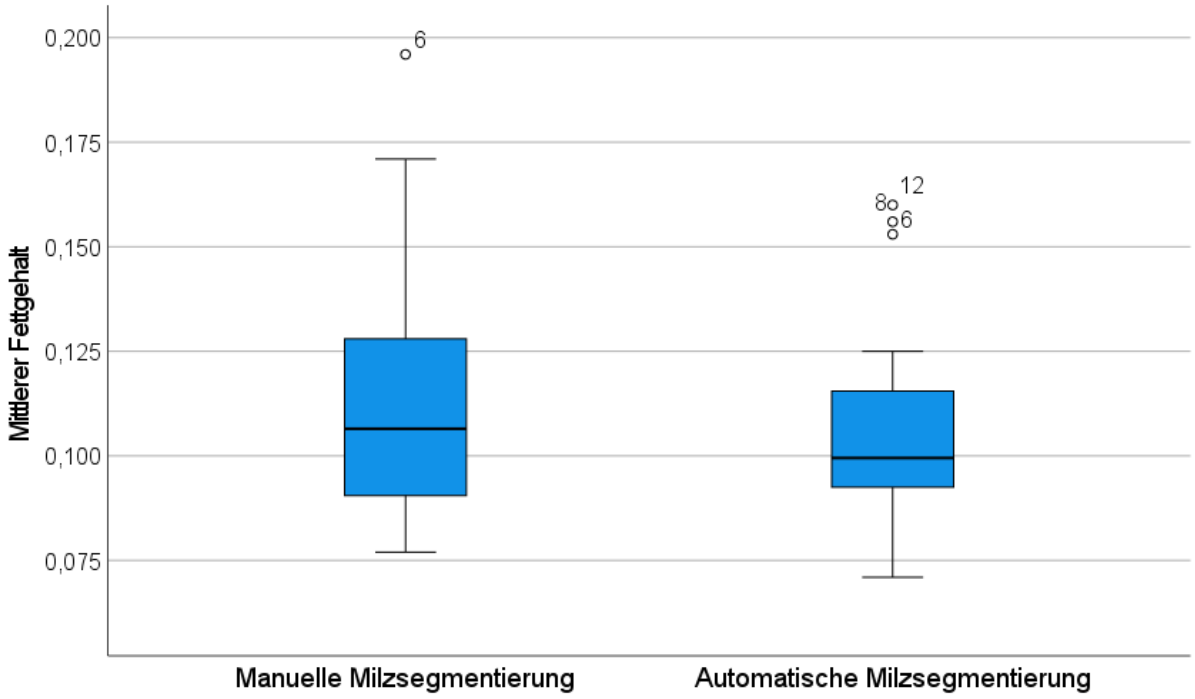

Abbildung 27, Vergleich Fettgehalt Milz

Es wurde die Korrelation des Organvolumens und des mittleren Fettgehaltes der manuellen und automatischen Organsegmentierung beider Organe, Leber und Milz, untersucht und auf Signifikanz untersucht.
In Tabelle 11 wurde die Korrelation des Lebervolumens untersucht. Es zeigte sich ein Korrelationskoeffizient von 0,967. Das Ergebnis ist hochsignifikant.

		Manuelle Lebersegmentierung	Automatische Lebersegmentierung
Manuelle Lebersegmentierung	Korrelationskoeffizient	1,000	0,967
	Sig. (1-seitig)		<.001
Automatische Lebersegmentierung	Korrelationskoeffizient	0,967	1,000
	Sig. (1-seitig)	<.001	

Tabelle 11, Korrelation nach Spearman-Rho Organvolumen Leber

Abbildung 28 zeigt eine graphische Darstellung des Organvolumens der Lebersegmentierungen. Man erkennt eine enge Anordnung der Werte um eine gezogene Anpassungslinie. Das Bestimmtheitsmaß R^2 der Graphik liegt bei 0,938 und unterstützt die berechnete Signifikanz.

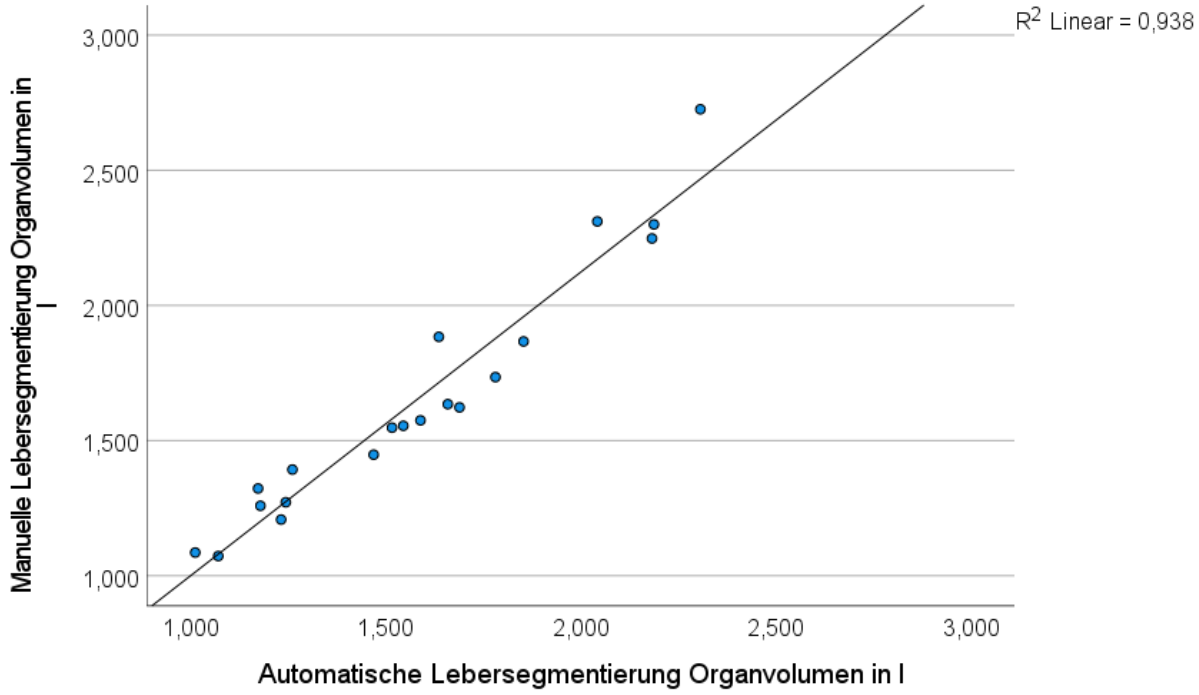

Abbildung 28, Graphische Darstellung Korrelation Organvolumen Lebersegmentierung

In Tabelle 12 wurde die Korrelation des Organvolumens für die Milz untersucht. Der Korrelationskoeffizient betrug 0,908 und das Ergebnis ist ebenfalls hochsignifikant.

		Manuelle Milzsegmentierung	Automatische Milzsegmentierung
Manuelle Milzsegmentierung	Korrelationskoeffizient	1,000	0,908
	Sig. (1-seitig)		<.001
Automatische Milzsegmentierung	Korrelationskoeffizient	0,908	1,000
	Sig. (1-seitig)	<.001	

Tabelle 12, Korrelation nach Spearman-Rho Organvolumen Milz

Abbildung 29 veranschaulicht den Vergleich der Organvolumina der Milzsegmentierungen. Man erkennt den in Abbildung 25 ebenfalls zu sehenden statistischen Ausreißer. Das Bestimmtheitsmaß R^2 der Graphik liegt bei 0,632 und unterstützt die berechnete Signifikanz.

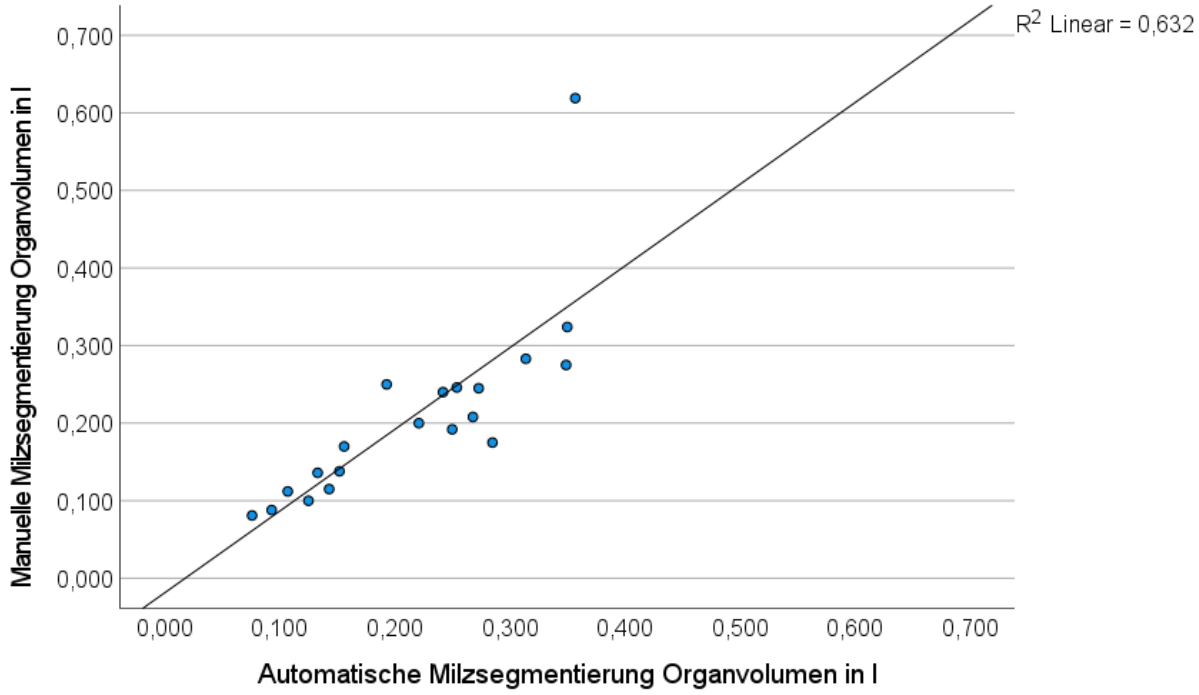

Abbildung 29, Graphische Darstellung Korrelation Organvolumen Milzsegmentierung

Die Untersuchung auf Signifikanz des mittleren Fettgehaltes der Lebersegmentierungen wurde in Tabelle 13 dargestellt. Der Korrelationskoeffizient beträgt 0,972, das Ergebnis ist signifikant.

		Manuelle Lebersegmentierung	Automatische Lebersegmentierung
Manuelle Lebersegmentierung	Korrelationskoeffizient	1,000	0,972
	Sig. (1-seitig)		<.001
Automatische Lebersegmentierung	Korrelationskoeffizient	0,972	1,000
	Sig. (1-seitig)	<.001	

Tabelle 13, Korrelation nach Spearman-Rho mittlerer Fettgehalt Leber

Abbildung 30 zeigt die graphische Darstellung der Korrelation des mittleren Fettgehaltes der Lebersegmentierungen. Man erkennt eine enge Anordnung der Werte um eine gezogene Anpassungslinie. Das Bestimmtheitsmaß R^2 der Graphik liegt bei 0,997 und unterstützt die berechnete Signifikanz.

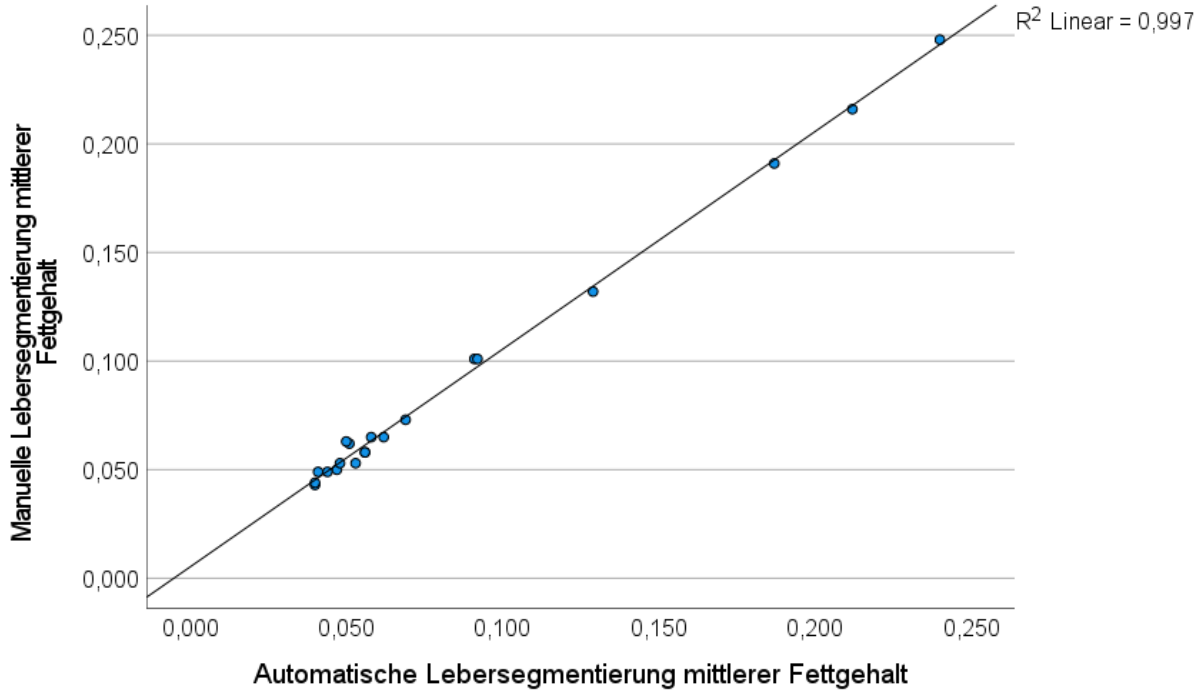

Abbildung 30, Graphische Darstellung Korrelation mittlerer Fettgehalt Leber

Die Untersuchung der Milzsegmentierungen auf Korrelation des mittleren Fettgehaltes wurde in Tabelle 14 dargestellt. Der Korrelationskoeffizient beträgt 0,823, das Ergebnis ist signifikant.

		Manuelle Milzsegmentierung	Automatische Milzsegmentierung
Manuelle Milzsegmentierung	Korrelationskoeffizient	1,000	0,823
	Sig. (1-seitig)		<.001
Automatische Milzsegmentierung	Korrelationskoeffizient	0,823	1,000
	Sig. (1-seitig)	<.001	

Tabelle 14, Korrelation nach Spearman-Rho mittlerer Fettgehalt Milz

Es erfolgte eine graphische Veranschaulichung in Abbildung 31. Man erkennt eine etwas lockerere Anordnung der Werte um eine gezogene Anpassungslinie als in Abbildung 30. Das Bestimmtheitsmaß R^2 der Graphik liegt bei 0,865 und unterstützt die berechnete Signifikanz.

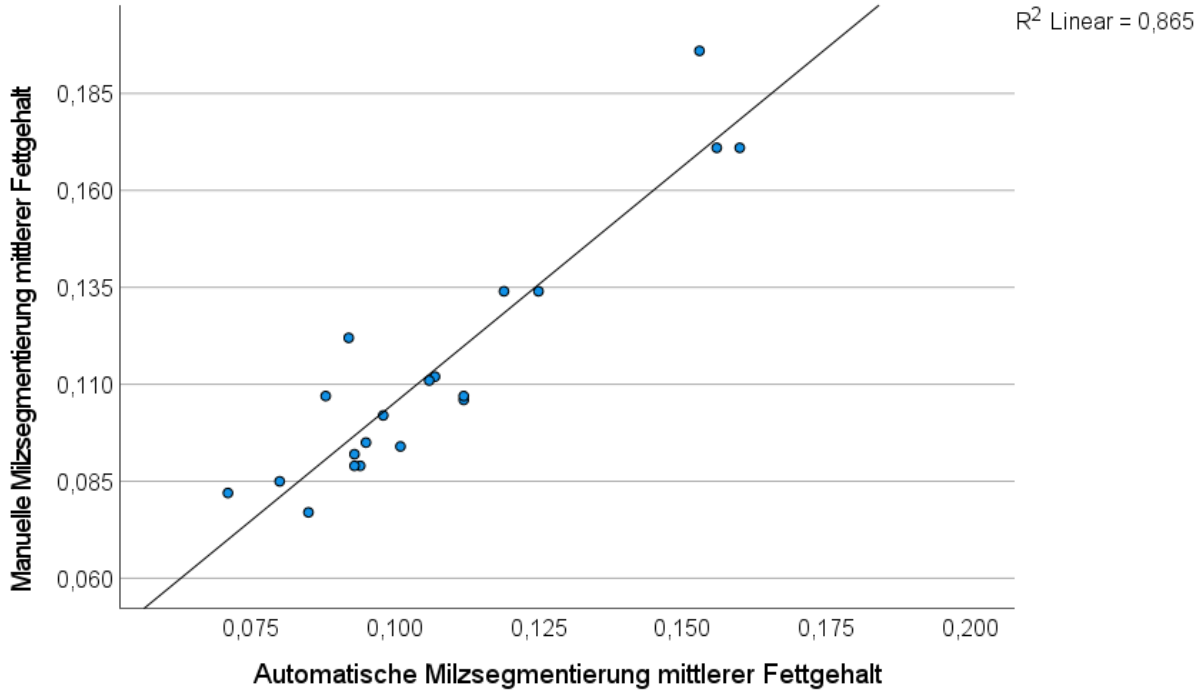

Abbildung 31, Graphische Darstellung Korrelation mittlerer Fettgehalt Milz

5. Diskussion

In dieser Arbeit wurde geprüft, wie ein Algorithmus mithilfe von maschinellem Lernen auf Grundlage eines CNN zur vollautomatischen Segmentierung von Leber und Milz etabliert und validiert werden kann. Ziel war es, ein Werkzeug zu erhalten, um Vorteile bei der Auswertung großer Mengen an MRT Aufnahmen im Rahmen der NAKO Kohortenstudie zu schaffen. Im Ergebnisteil konnte eindrücklich gezeigt werden, dass für alle untersuchten Parameter die maschinelle Auswertung mit der manuellen gleichzusetzen ist und dabei weniger Zeit benötigt. Befürchtete Organverwechslungen, welche für großangelegte Untersuchungen unbedingt zu vermeiden sind, wurden nicht festgestellt.

Es zeigte sich, dass die erhobenen Ergebnisse im Wesentlichen mit denen gängiger Literatur zum Thema automatische Organsegmentierung auf Grundlage von CNN übereinstimmten.

Wie bereits durch Lavdas und Glocker beschrieben, gilt dieses Verfahren der automatisierten, Algorithmus gestützten Auswertung als zeitsparendes und präzises Tool um eine automatische Erfassung verschiedener parenchymatöser Organe aus MRT Daten zu generieren (19).

Es hat sich gezeigt, dass die vollautomatische Organsegmentierung mittels CNN Algorithmus effektiv und mit hoher Genauigkeit möglich ist. Insbesondere bei der automatischen Segmentierung der Leber ließen sich hohe Werte für Präzision ermitteln. Der verwendete Algorithmus war imstande das Organ zu erkennen und es konnte eine hohe Genauigkeit der Segmentation bezüglich der Einhaltung der Organgrenzen beobachtet werden. Für das Organ Leber ergaben sich hohe Übereinstimmungen, bei den annähernd vergleichbaren Fettgehalt und Organvolumina kann dies als plausibel validiert werden. In Vorarbeiten wurden CNN Algorithmen für das Organ Leber bereits benutzt. So zeigte sich bei Wang und Song eine Präzision von 88,3%±SD bei Verwendung eines CNN zur Qualitätsbewertung von Leber MRT Aufnahmen (46). Auch in der Arbeit von Azer zeigte sich generell eine gute Leistung von CNN, im Bereich der Segmentierung, aber auch der Läsionsdetektion (47). Die Ausarbeitungen von Lu und Wu (48), sowie Hu und Wu (49) basieren zwar auf computertomographischen (CT) Aufnahmen, erzielten bei der Segmentierung von der Leber mittels CNN Algorithmus eine hohe Korrelation zwischen manueller und automatischer Segmentierung. So ermittelten Lu und Wu bei der Lebersegmentierung

einen Dice Wert von 96,0%±SD und damit höher als hier errechnete Koeffizient von 93,9%±SD (48).

Für die Milz ließen sich ebenfalls gute Ergebnisse erzielen. Es zeigte sich, dass bei der automatischen Segmentierung öfters organfremde Schichten segmentiert wurden, das Organ also in einigen Fällen nur partiell erkannt wurde. Die Validierung konnte mittels den Koeffizienten Dice und Jaccard nachgewiesen werden. Des Weiteren korrelierten Organvolumina der manuellen und automatischen Segmentierung der Milz miteinander. Die Sekundärparameter Volumen und Fettgehalt wurden ebenfalls mit gutem Ergebnis verglichen. Bei der Milzsegmentierung wurden nur geringfügig niedrigere Werte ermittelt. Diese Ergebnisse sprechen ebenfalls für eine gute Übereinstimmung, welches sich ebenfalls beim Vergleich von Fettgehalt und Organvolumen gezeigt hat. Die Milz hat einen generell niedrigen Fettanteil. Das bedeutet, dass fetthaltiges Gewebe aus dem Milzhilus je nach Segmentierung eingeschlossen wurde. Dieser mögliche Umstand schmälert die Aussagekraft des Sekundärparameters Fettgehalt.

Im Vergleich zu den Ergebnissen für das Organ Leber gab es Abzüge unter anderem im Bereich der Präzision. In der Arbeit von Hu und Wu ergab sich bei der automatischen Milzsegmentierung ein Dice Wert von 94,2%±SD, und somit höher gelegen als der hier ermittelte durchschnittliche Wert von 74,2%±SD. Anzumerken gilt, dass die zugrunde liegenden Daten auf CT Aufnahmen basierten (49). Lavdas und Glocker konnten ebenfalls hohe Übereinstimmungen bei ihrer Arbeit zu KI unterstützter Multiorgansegmentierung im MRT erzielen (19).

Die Vorarbeiten zielen gehäuft auf das Organ Leber ab, auch mit der besonderen Bedeutung der Leber im Vergleich zum Organ Milz. Im Fokus dieser Arbeiten stand besonders die Läsionsdetektion, die nicht Gegenstand vorliegender Dissertation war. Grund für die mit der aktuellen Forschung nicht im Einklang stehende Arbeit ist die Fragestellung nach vollautomatischer Segmentationsmöglichkeit im Rahmen der NAKO Studie.

Die Ergebnisse dieser Arbeit für das Organ Leber spiegeln auch die Ergebnisse aktueller Forschung wider. Für das Organ Milz ergaben sich Diskrepanzen beispielsweise beim errechneten Dice Wert von 74,2%±SD im Vergleich zur Arbeit von Hu und Wu mit einem Dice Wert von 94,2%±SD (49). Es zeigte sich, dass besonders beim Organ Milz Muskulatur im Bereich der linken Schulter fehlsegmentiert worden ist.

Dies kann eventuell an vergleichbarem Erscheinungsbild der fehlsegmentierten Struktur liegen. Erklärungsansatz bietet eventuell der vergleichbare Bildkontrast, der zur Fehlklassifizierung führte oder auch die entfernt ähnliche Form des M. Deltoideus, halbmondförmig konfiguriert. Diesem Problem ließe sich begegnen, indem man dem Algorithmus beispielsweise Regionen vorgibt, in denen das Organ zu segmentieren ist, sodass der CNN Algorithmus bei der Aufgabe die Milz zu segmentieren ausschließlich in der Bauchhöhle arbeitet.

Unter anderem konnten auch vereinzelt rechteckige Lücken in der Organdarstellung gezeigt werden. Am ehesten könnte sich das mit der Rasterform des CNN Filters begründen, welcher in gewisser Reihenfolge über jedes einzelne Bild bewegt wird.

Es existieren bereits etliche etablierte Machine Learning Algorithmen, die auch als Alternativen zu CNN zu sehen sind, wie zum Beispiel Classification Forests oder Multi-Atlas Approaches (19). In einer Vorarbeit hat sich herauskristallisiert, dass der Einsatz eines CNN aufgrund der hohen Präzision den anderen Algorithmen vorzuziehen ist (19).

Es konnte schon für die Kohortenstudie KORA durch MRT Aufnahmen und mittels Implementierung eines CNN Algorithmus, mit gleichen Hyperparametern wie in diesem Fall, gezeigt werden, dass dies eine effektive und günstige Methode zur Organsegmentierung darstellt (20). Andere Arbeiten versuchten bereits mittels Machine Learning Pathologien in MRT Aufnahmen automatisch zu detektieren. Beispielsweise wurden Ansätze verfolgt, um Leberfibrosen zu klassifizieren (50). Auch in der Tumorsuche fanden CNN Algorithmen Anwendung und wurden implementiert (51, 52).

Des Weiteren, nur um vereinzelte Ausblicke und vorangegangene Arbeiten zu nennen, sind bereits verschiedenste Organe mittels CNN Algorithmen segmentiert worden, zum Beispiel Prostata (53) oder Pankreas (54). Jedoch lassen sich die Einsatzgebiete der CNN Algorithmen nur schwer begrenzen. Einzelne Limitationen stellen am ehesten schwer differenzierbare Organgrenzen, bzw. Homogenitäten dar, aufgrund des meistens kantenbasierten Segmentierungsverfahrens.

Bei Kohortenstudien fallen große Mengen an Daten an, sogenannte *Big Data Studien* mit erheblichen Teilnehmerzahlen. Durch Steigerung der Verfügbarkeit von MRT Geräten und der immer höheren Einsatzbereitschaft auch in Studien, fallen letztlich

eine Menge an Rohdaten und auszuwertenden MRT Aufnahmen an, welche durch den einzelnen Menschen manuell nicht mehr zu bewältigen sind. Hier kommen *Machine Learning* Algorithmen zum Tragen, um Werte wie Organvolumina oder Fettgehalt automatisch zu bestimmen. Insbesondere für Organe wie die Leber können das Organvolumen und der Fettgehalt von großer Wichtigkeit für die Prädiktion und Prognosen von Krankheiten sein. Kommt es zu Follow-up Untersuchungen werden Vergleiche mit diesen Daten in Kombination mit anderen erhobenen Werten beispielsweise zu wertvollen Erkenntnissen für die Entstehung von sogenannten Volkskrankheiten, wie Diabetes oder Leberverfettung, führen. Mittels des einmal etablierten und validierten CNN Algorithmus lassen sich jedwede MRT Aufnahmen auch im klinischen Alltag analysieren. Somit steht durch relativ mäßigen Aufwand zu Beginn der Implementierung am Ende ein wertvolles und langfristig einsetzbares Werkzeug zur Verfügung.

Mit mäßigem Aufwand in Hinsicht auf technische Umsetzung, Erstellung von Trainingsdatensätzen, und nötiger technischer Ausstattung, lassen sich Algorithmen zur vollautomatischen Organsegmentierung erstellen, die wie in diesem Fall mit besonders hoher Genauigkeit arbeiten. Besonders in Hinblick auf den Einsatz im Bildgebungsmaterial der Kohortenstudie NAKO und bei eventuellen späteren Einsätzen im klinischen Alltag ergeben sich enorme Vorteile.

Die im Vergleich zu Vorarbeiten schlechteren Ergebnisse für das Organ Milz schränken die Aussagekraft des segmentierenden Algorithmus ein. Im Vergleich zur manuellen Segmentierung ergaben sich die oben beschriebenen Diskrepanzen. Dies macht eine Nachbesserung des Algorithmus für die Milzsegmentierung unabdingbar. Gerade im Bereich Organerkennung und Präzision sollten noch weitere Anpassungen erfolgen.

Limitationen der Arbeit sind unter anderem die Erstellung von Trainingsdatensätzen. Die Trainingsprozedur, und somit auch der Trainingserfolg, hängt in hohem Maße von der Expertise des Erstellers ab. So könnten sich eventuell bei anfänglicher Segmentierung durch erfahreneres Personal auch bessere Werte für Präzision ergeben.

Sind die zugrundeliegenden Daten, und damit die Trainingsdaten, begrenzt, muss man eventuell auf einen anderen Algorithmus als CNN zurückgreifen. Des Weiteren ist der implementierte Algorithmus spezifisch für MRT Aufnahmen und müsste bei CT

Aufnahmen mit neuen Trainingsdatensätzen gespeist werden. Nebst der Hardware für die Erstellung und Berechnung, muss auch das notwendige, geschulte, Personal zur Verfügung stehen, um die Grundsegmentierungen, aber auch die Datenverarbeitung vorzunehmen. Weitere Nachteile sind beispielsweise die Notwendigkeit jede Struktur, die segmentiert werden soll, vorher auch manuell in entsprechender Anzahl zu Trainingszwecken zu segmentieren. Andere, im MRT schlechter abgrenzbare Organe bzw. Strukturen als Leber und Milz, liefern eventuell schlechtere Resultate, oder lassen sich mit geringerem Erfolg automatisch segmentieren. Die hier gewählten Organe sind durch ihre Konfiguration visuell gut abgrenzbar und somit auch für den CNN Algorithmus gut zu differenzieren. Sind die Organgrenzen weniger Kontraststark oder inhomogener könnten sich Probleme ergeben.

Im Umkehrschluss gilt anzumerken, dass voll implementierte und validierte Methoden zur automatischen Organsegmentierung auf unbestimmte Zeit erhebliche Möglichkeiten offerieren. Einmal mittels manuellen Trainingsdaten etabliert, lassen sich die Algorithmen in vielfältigen Fällen im klinischen Alltag einsetzen. Die Möglichkeiten reichen von direkt mit der Bildgebung ausgewertete Organgrößen, bis hin zur vollautomatischen Erkennung und Hervorhebung von Läsionen in der Standard – oder Notfallbildgebung.

Es ist eher als unwahrscheinlich anzusehen, dass, bei heutigem Stand von Technik, der menschliche Faktor gänzlich aus der medizinischen Bildanalyse wegzudenken ist. Es konnte gezeigt werden, dass die automatische Segmentierung dennoch einer gewissen Aufsicht bzw. Nachbearbeitung bedarf. Gerade bei auftretenden Fehlsegmentierungen, in denen außerhalb der Organgrenzen segmentiert wurde, muss eine Korrektur nach der stattgefundenen Segmentierung erfolgen.

Technisch gesehen, ließe sich dies durch zusätzliche Informationsdarbietung beheben, das heißt beispielsweise durch vorherige Festlegung der zu erwartenden Organposition im Situs.

Plausible Einsatzmöglichkeiten der automatischen Bildanalyse bieten sich für die Telemedizin. Es kann eine automatische Bildanalyse ohne beiwohnenden Radiologen erfolgen, detektiert der Algorithmus eine Auffälligkeit, wird ein Arzt darauf hingewiesen.

Rasanter Fortschritt im Bereich der KI, scheinbar grenzenloses Potenzial, die Möglichkeit für Jedermann damit zu arbeiten sind nur geringe Anlässe, die Grund zur

Auseinandersetzung mit diesem Thema bieten sollten. Nicht nur im Bereich des CNN mit der automatischen Bildanalyse, sondern auch ganz allgemein im Bereich des KI Einsatzes in der Humanmedizin sollten unbedingt Regelkodexe ermittelt, Richtlinien und Grenzen festgelegt werden. Läuft man doch Gefahr, eines Tages nicht mehr Teil der Entscheidungskette zu sein.

Für die MRT Aufnahmen, die aus der NAKO Studie gewonnen wurden, bedeutet dieser etablierte CNN Algorithmus einen enormen Vorteil für die medizinische Bildanalyse und eine zeitnahe Bearbeitung.

“We must address, individually and collectively, moral and ethical issues raised by cutting-edge research in artificial intelligence and biotechnology, which will enable significant life extension, designer babies, and memory extraction.” —Klaus Schwab (55)

6. Zusammenfassung

Die vorliegende Dissertation widmet sich der Evaluierung von Deep-Learning-Methoden für die automatische Segmentierung von Leber und Milz in MRT-Bildern, basierend auf der Nationalen Kohorte (NAKO), einer Langzeitstudie mit dem Ziel, Gesundheitsdaten von rund 200.000 Personen über einen Zeitraum von bis zu 30 Jahren zu erfassen. Ein besonderer Fokus dieser Studie liegt auf der MRT-Bildgebung von 30.000 Teilnehmern zur Identifikation von Faktoren, die zur Entstehung von Volkskrankheiten beitragen.

Angesichts der Herausforderung, dass die manuelle Analyse von medizinischen Bildern zeitintensiv ist, wurde in dieser Arbeit ein Convolutional Neural Network (CNN) zur Unterstützung der Bildanalyse eingesetzt. Für die Studie wurden 100 MRT-Aufnahmen ausgewählt, in denen Leber und Milz manuell segmentiert wurden, um ein Trainingsset für das CNN zu erstellen. 80 dieser Aufnahmen dienten als Trainingsdaten, während die verbleibenden 20 Aufnahmen genutzt wurden, um die Leistungsfähigkeit des CNNs zu evaluieren, indem ihre automatische Segmentierung mit den manuellen Segmentierungen verglichen wurde.

Die Ergebnisse der Dissertation zeigen, dass CNNs in der Lage sind, die Organe Leber und Milz mit hoher Präzision zu segmentieren. Diese Präzision wurde durch die Analyse von Übereinstimmungen, Organvolumina und Fettgehalt bestätigt. Weiterhin wurde die Genauigkeit der automatischen Segmentierung mittels statistischer Methoden, wie den Dice- und Jaccard-Koeffizienten, verifiziert, die signifikante Korrelationen zu den manuellen Segmentierungen aufzeigten.

Zusammengefasst verdeutlicht diese Arbeit das Potenzial von Deep-Learning-Algorithmen, um die Effizienz und Genauigkeit in der medizinischen Bildanalyse zu steigern. Der Einsatz dieser Technologien könnte zukünftig nicht nur erhebliche Zeitersparnisse ermöglichen, sondern auch den Weg zu einer verbesserten und möglicherweise vollautomatischen medizinischen Diagnostik ebnen.

7. Literaturverzeichnis

1. Röhrig BP, Jean-Baptist du; Blettner, Maria. Studiendesign in der medizinischen Forschung. Dtsch Arztebl Int 2009. 2009;11/2009(106(11): 184-9).
2. Röhrig BP, Jean-Baptist du; Wachtlin, Daniel; Blettner, Maria. Studientypen in der medizinischen Forschung. Dtsch Arztebl Int 2009. 2009;15/2009(106(15): 262-8).
3. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet. 2002;359(9303):341-5.
4. German National Cohort C. The German National Cohort: aims, study design and organization. Eur J Epidemiol. 2014;29(5):371-82.
5. Bamberg F, Kauczor HU, Weckbach S, Schlett CL, Forsting M, Ladd SC, et al. Whole-Body MR Imaging in the German National Cohort: Rationale, Design, and Technical Background. Radiology. 2015;277(1):206-20.
6. Belbasis L, Bellou V. Introduction to Epidemiological Studies. Methods Mol Biol. 2018;1793:1-6.
7. Ahrens W, Jockel KH. [The benefit of large-scale cohort studies for health research: the example of the German National Cohort]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2015;58(8):813-21.
8. NAKO Deutschlandkarte [Available from: https://nako.de/studienteilnehmer/studienzentren/.
9. Schlett CL, Hendel T, Weckbach S, Reiser M, Kauczor HU, Nikolaou K, et al. Population-Based Imaging and Radiomics: Rationale and Perspective of the German National Cohort MRI Study. Rofo. 2016;188(7):652-61.
10. Dössel O. Bildgebende Verfahren in der Medizin. 2 ed: Springer Vieweg, Berlin, Heidelberg; 2016. XXX, 513 p.
11. Dominik Weishaupt VDK, Borut Marincek. Eine Einführung in Physik und Funktionsweise der Magnetresonanzbildgebung: Springer-Verlag Berlin Heidelberg; 2009.
12. Heinz Handels TMD, Andreas Maier, Klaus H. Maier-Hein, Christoph Palm, Thomas Tolxdorff. Medizinische Bildverarbeitung. 2 ed: Vieweg+Teubner; 2009. XVI, 432 p.
13. Iserhardt-Bauer S. Standardisierte Protokolle für die medizinische Bildanalyse und Bildvisualisierung: Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart

2010.
14. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. Med Phys. 2014;41(5):050902.
15. Kavur AE, Gezer NS, Baris M, Sahin Y, Ozkan S, Baydar B, et al. Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors. Diagn Interv Radiol. 2020;26(1):11-21.
16. Sankur B. Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging. 2004;13(1).
17. Kaganami HG, & Beiji, Z. Region-Based Segmentation versus Edge Detection. 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing2009.
18. Image Segmentation in Deep Learning: Methods and Applications: missinglink.ai; [Available from: https://missinglink.ai/guides/computer-vision/image-segmentation-deep-learning-methods-applications/.
19. Lavdas I, Glocker B, Kamnitsas K, Rueckert D, Mair H, Sandhu A, et al. Fully automatic, multiorgan segmentation in normal whole body magnetic resonance imaging (MRI), using classification forests (CFs), convolutional neural networks (CNNs), and a multi-atlas (MA) approach. Med Phys. 2017;44(10):5210-20.
20. Thomas Küstner SM, Marc Fischer, Jakob Weibeta, Konstantin Nikolaou, Fabian Bamberg, Bin Yang, Fritz Schick, Sergios Gatidis. SEMANTIC ORGAN SEGMENTATION IN 3D WHOLE-BODY MR IMAGES. ICIP 2018: 3498-3502. 2018.

21. Reig B, Heacock L, Geras KJ, Moy L. Machine learning in breast MRI. J Magn Reson Imaging. 2019.
22. Baris Kayalibay GJ, and Patrick van der, Smagt. Cnn-based segmentation of medical imaging data. CoRR. 2017;Vol. abs/1701.03056.
23. Zisserman KSaA. Very deep convolutional networks for large-scale image recognition. ICLR 20152015.
24. A. Krizhevsky IS, G. E. Hinton, editor ImageNet Classification with Deep Convolutional Neural Networks. NIPS; 2012.
25. Olaf Ronneberger PF, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015,:pp. 234–41.
26. F. Milletari NN, and S.-A. Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. ArXiv e-prints. 2016;June 2016.
27. Kamnitsas KaL, Christian and Newcombe, Virginia F.J. and Simpson, Joanna P. and Kane, Andrew D. and Menon, David K. and Rueckert, Daniel and Glocker, Ben. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis, Elsevier BV. 2017;36:61–78.
28. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017;19:221-48.
29. Krizhevsky A SI, Hinton GE. ImageNet classification with deep convolutional neural networks. NIPS'12 Proceedings of the 25thInternational Conference on Neural Information Processing Systems. 2012;Volume 1:p 1097–105.
30. Kim S. toward data sience. 2019.
31. Brownlee J. A Gentle Introduction to Learning Curves for Diagnosing Machine Learning Model Performance 2019 [Available from: https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/.
32. Ian Goodfellow YB, Aaron Courville. Deep Learning : das umfassende Handbuch : Grundlagen, aktuelle Verfahren und Algorithmen, neue Forschungsansätze: Frechen: MITP, 2018.
33. Bengio YJaDKaHMaS. Predicting the Generalization Gap in Deep Networks with Margin Distributions2019. Available from: https://arxiv.org/pdf/1810.00113.pdf.
34. Merkert P. Maschinelle Übersetzer: DeepL macht Google Translate Konkurrenz2017. Available from: https://www.heise.de/newsticker/meldung/Maschinelle-Uebersetzer-DeepL-macht-Google-Translate-Konkurrenz-3813882.html.
35. Gatidis S, Heber SD, Storz C, Bamberg F. Population-based imaging biobanks as source of big data. Radiol Med. 2017;122(6):430-6.
36. Dixon WT. Simple proton spectroscopic imaging. Radiology. 1984;153(1):189-94.
37. MicroDicom DICOM Viewer. Copyright © 2007-2019 MicroDicom.
38. Wolf I, Vetter M, Wegner I, Bottger T, Nolden M, Schobinger M, et al. The medical imaging interaction toolkit. Med Image Anal. 2005;9(6):594-604.
39. The Medical Imaging Interaction Toolkit (MITK). German Cancer Research Center, Division of Medical Image Computing, Im Neuenheimer Feld 280, 69120 Heidelberg.
40. Kaiming He XZ, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016,:pp. 770–8.
41. Gao Huang ZL, Kilian QWeinberger, and Laurens, Maaten vd. Densely connected convolutional networks. arXiv preprint arXiv:160806993. 2016.
42. Aliza Medical Imaging & DICOM Viewer. Copyright (c) 2014-2019 Aliza Medical Imaging, Bonn, Germany.
43. Dateiformat .nii [Available from: https://nifti.nimh.nih.gov/nifti-1/.
44. Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimization for Medical Image Segmentation: Theory and Practice When Evaluating With Dice Score or Jaccard Index. IEEE Trans Med Imaging. 2020;39(11):3679-90.
45. Ma J. Dixon techniques for water and fat imaging. J Magn Reson Imaging. 2008;28(3):543-58.

46. Wang Y, Song Y, Wang F, Sun J, Gao X, Han Z, et al. A two-step automated quality assessment for liver MR images based on convolutional neural network. Eur J Radiol. 2020;124:108822.
47. Azer SA. Deep learning with convolutional neural networks for identification of liver masses and hepatocellular carcinoma: A systematic review. World J Gastrointest Oncol. 2019;11(12):1218-30.
48. Lu F, Wu F, Hu P, Peng Z, Kong D. Automatic 3D liver location and segmentation via convolutional neural network and graph cut. Int J Comput Assist Radiol Surg. 2017;12(2):171-82.
49. Hu P, Wu F, Peng J, Bao Y, Chen F, Kong D. Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. Int J Comput Assist Radiol Surg. 2017;12(3):399-411.
50. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver Fibrosis: Deep Convolutional Neural Network for Staging by Using Gadoxetic Acid-enhanced Hepatobiliary Phase MR Images. Radiology. 2018;287(1):146-55.
51. Iqbal S, Ghani MU, Saba T, Rehman A. Brain tumor segmentation in multi-spectral MRI using convolutional neural networks (CNN). Microsc Res Tech. 2018;81(4):419-27.
52. Pereira S, Pinto A, Alves V, Silva CA. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. IEEE Trans Med Imaging. 2016;35(5):1240-51.
53. Clark T, Zhang J, Baig S, Wong A, Haider MA, Khalvati F. Fully automated segmentation of prostate whole gland and transition zone in diffusion-weighted MRI using convolutional neural networks. J Med Imaging (Bellingham). 2017;4(4):041307.
54. Cai J, Lu L, Zhang Z, Xing F, Yang L, Yin Q. Pancreas Segmentation in MRI using Graph-Based Decision Fusion on Convolutional Neural Networks. Med Image Comput Comput Assist Interv. 2016;9901:442-50.
55. Marr B. 28 Best Quotes About Artificial Intelligence2017. Available from: https://www.forbes.com/sites/bernardmarr/2017/07/25/28-best-quotes-about-artificial-intelligence/#58c5d6e14a6f.

8. Erklärung zum Eigenanteil

Die Arbeit wurde in der Radiologischen Universitätsklinik Tübingen in der Abteilung für Diagnostische und Interventionelle Radiologie unter der Betreuung von PD Dr. Sergios Gatidis durchgeführt.

Die Konzeption der Studie erfolgte durch PD Dr. Sergios Gatidis. Er hat die Arbeit betreut und die Erstellung des Manuskripts begleitet.

Die MRT-Aufnahmen sowie die dazugehörigen demographischen Daten wurden von der NAKO Gesundheitsstudie zur Verfügung gestellt.

Die Entwicklung, Implementierung und der Trainings- sowie Inferenzprozess des maschinellen Lernalgorithmus mithilfe der Daten der NAKO-Studie sowie die automatische Evaluation erfolgten durch Sarah Müller, ebenfalls die Berechnung der Metriken unter den Punkten 4.3.2.2 und 4.3.2.3, sowie die Erstellung der Abbildung 12 im Ergebnisteil. Die Datengewinnung und Auswertung unter Punkt 4.2.1 erfolgte in Zusammenarbeit mit Sarah Müller.

Die Segmentierungsanalysen wurden von Mitarbeitern der Abteilung für Diagnostische und Interventionelle Radiologie am Universitätsklinikum Tübingen durchgeführt.

Die Bestimmung des Fettgehalts und der Organvolumina aus den manuell und automatisch erstellten Masken, wie auch die Erstellung der Abbildungen 10 und 11 erfolgten durch Dr. Tobias Hepp.

Die statistische Auswertung erfolgte mit Unterstützung durch PD Dr. Sergios Gatidis und Dr. Tobias Hepp.

Ich versichere das Manuskript selbstständig verfasst zu haben und keine weiteren als die von mir angegebenen Quellen verwendet zu haben.

Tübingen, 13.11.2024, Lukas Fetzer

9. Danksagung

Zunächst möchte ich mich bei meinem Doktorvater, PD Dr. Sergios Gatidis bedanken. Für seine unschätzbare Unterstützung und herausragende Geduld.

Des weiteren gilt mein Dank Sarah Müller für die Durchführung des CNN Algorithmus und die ausführliche Hilfe bei technischen Fragen.

Dr. Tobias Hepp möchte ich für die Unterstützung bei der statistischen Auswertung danken.

Außerdem allen weiteren Unterstützern und Korrekturlesern.