Technische Universität München
TUM School of Medicine and Health

# Entwicklung von 2D- und 3D-Deep-Learning-Ansätzen zur Differenzierung von Atypischen Lipomatösen Tumoren von Lipomen basierend auf magnetresonanztomografischen Bildern

Daniel Wolfgang Kramp

Vollständiger Abdruck der von der TUM School of Medicine and Health zur Erlangung eines Doktors der Medizin (Dr. med.) genehmigten Dissertation.

Vorsitz: Prof. Kathrin Schumann, Ph.D.

Prüfende der Dissertation:

1. Priv.-Doz. Dr. Benedikt Schwaiger
2. Priv.-Doz. Dr. Jan Peeken

Die Dissertation wurde am 22.03.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Medicine and Health am 07.08.2024 angenommen.

# Inhaltsverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| AI..... | Artificial Intelligence |
| MRT..... | Magnetresonanztomografie |
| ALT..... | Atypischer lipomatöser Tumor |
| CT..... | Computertomografie |
| FISH..... | Fluoreszenz-in-situ-Hybridisierung |
| MDM2..... | Mouse double minute 2 |
| WHO..... | World Health Organization |
| T1w..... | T1-gewichtete Spin-Echo-Sequenz |
| T2w..... | T2-gewichtete Spin-Echo-Sequenz |
| T1fsgd…….... | Kontrastverstärkte T1-gewichtete fettgesättigte Sequenzen |
| NIfTI..... | Neuroimaging Informatics Technology Initiative |
| 2D..... | Zweidimensional |
| 3D..... | Dreidimensional |
| ROI ..... | Region of Interest |
| DL..... | Deep Learning |
| CAM..... | Class Activation Map |
| CNN..... | Convolutional Neural Network |
| LiDAR..... | Light Detection and Ranging |

# 1. Einleitung

## 1.1 Lipome und atypische lipomatöse Tumoren

Bufalini beschrieb im Jahre 1925 die röntgenologischen Eigenschaften von Lipomen und charakterisierte diese als einen Bereich im Körper, der erhöhte Röntgendurchlässigkeit besitzt (Bufalini-Zeichen) (Kindblom et al., 1974). Im darauffolgenden Jahr wurden die tiefergelegenen Lipome von Moriconi als entweder inter- oder intramuskulär beschrieben. Die frühzeitige Differenzierung von Lipomen und Liposarkomen ist dabei entscheidend, die zum Teil zu unnötigen Amputationen bei Lipomen und erhöhter Morbidität durch unsachgemäße Entfernung von Sarkomen führen (Tedesco & Henshaw, 2016).
Die Schwierigkeit in der Unterscheidung beruht auf Ähnlichkeiten im Aussehen von Lipomen und gut differenzierten lipomatösen Tumoren mit weniger gut differenzierten lipomatösen Tumoren in der Bildgebung.

In vorangehenden Studien wurden bereits bildgebende Merkmale bei der Diagnostik von Liposarkomen untersucht. Ein Hinweis auf ein Liposarkom besteht, wenn die Läsion einen maximalen Durchmesser von mehr als 10 cm aufweist ($P < .001$), kräftige Septen hat ($P = .001$), knotige Veränderungen und nicht adipöse Anteile beinhaltet, ($P = .003$) sowie weniger als 75 % Fettanteil aufweist ($P < .001$) (Kransdorf et al., 2002). Es wurde bereits gezeigt, dass in der MR-Bildgebung hohe Sensitivitäten und Spezifitäten durch die Fokussierung auf bestimmte Charakteristika wie Tumordurchmesser, Septendicke und Kontrastmittelaufnahme sowie die Lokalisation des Tumors zu erreichen sind (Knebel et al., 2019). Sowohl das Well-Differentiated Liposarcoma (WDL) als auch die ALTs sind nach der WHO-Klassifikation gut differenzierte Weichteiltumoren (G1) mit reifen Adipozyten, die nach der 2002 erschienenen „WHO Classification of Tumors“ sich im Hinblick auf die chirurgische Resizierbarkeit unterscheiden. Während gut differenzierte Liposarkome (WDLs) beispielsweise retroperitoneal gelegen sind und daher nicht vollständig reseziert werden können, so sind ALTs peripherer gelegen und können meist vollständig reseziert werden.

Die heute vor allem zur Differenzierung von Lipomen und Liposarkomen angewandte Diagnostik konzentriert sich vor allem auf die MRT bleibt weiterhin herausfordernd für die Radiolog*innen und erfordert viel bildgebende Erfahrung im Bereich der Weichteiltumore (Nagano et al., 2015; O'Donnell et al., 2013; Ryan et al., 2018). Trotz der Ähnlichkeit in der Bildgebung unterscheiden sich die Therapien je nach Weichteiltumorentität und Grading stark, weswegen eine genaue Klassifikation wichtig ist. Von den nicht-invasiven Maßnahmen ist die MRT, aufgrund des hohen Kontrastes der Weichgewebe, das Bildgebungsverfahren der ersten Wahl bei der Detektion und Differenzierung von Weichteiltumoren (Wu & Hochman, 2009).

Eine weit invasivere Möglichkeit stellt die histologische Untersuchung dar, sie ermöglicht jedoch auch die genauste pathologische Differenzierung. Hierfür werden sowohl histologische Merkmale als auch molekulare Merkmale genutzt. Um bei lipomatösen Neoplasien benigne und maligne lipomatöse Tumoren zu differenzieren, wird die molekulare immunhistochemische Analyse genutzt. Zielstruktur ist dabei das MDM2-Gen, das mithilfe einer Fluoreszenz-in-situ-Hybridisierung (FISH)-Analyse auf Amplifikation untersucht wird. Der MDM2-Amplifikations-Status ist bei ALTs (G1) erhöht, während der Status bei Lipomen im Normalbereich liegt (Jo & Fletcher, 2014; Moll & Petrenko, 2003).

Im Falle einer Amplifikation hemmt das MDM2-Gen das p53-Tumorsuppressor-Gen. Es stellt somit ein Protoonkogen dar (Moll & Petrenko, 2003).

Bei stark beanspruchten Zellen wird p53 phosphoryliert und kann schlecht von MDM2 gebunden werden. Es wird somit bei normaler Expression von MDM2 nur langsam abgebaut. Das hat zur Folge, dass sich die Expression des p53-Gens erhöht, die Zelle kann sich regenerieren und die Apoptose wird initiiert (Moll & Petrenko, 2003).
Lipomatöse Neoplasien sind die am häufigsten vorkommenden Tumoren, mit denen sich Ärzte beim erwachsenen Menschen auseinandersetzen. Der Häufigkeitsgipfel liegt zwischen dem 40. und dem 70. Lebensjahr. Bei adipösen Patienten ist die Auftretenswahrscheinlichkeit höher (Murphey et al., 2004; Myhre-Jen-

sen, 1981). Aufgrund des häufigen Ausbleibens von Symptomen wird die Inzidenz von lipomatösen Neoplasien vermutlich unterschätzt. Symptomlosigkeit führt dazu, dass viele Fälle nicht von Ärzten begutachtet und dokumentiert werden. In den Extremitäten sind proximal gelegene lipomatöse Tumoren häufiger maligne als distal gelegene, letzere sind häufiger benigne. Lipomatöse Neoplasien können sowohl oberflächlich als auch in der Tiefe auftreten und sind bei Männern etwas häufiger als bei Frauen anzutreffen. Kinder hingegen sind nur sehr selten betroffen. Am häufigsten treten subkutane Tumoren auf (Gassert et al., 2021; Goldblum et al., 2014).

Für die Differenzierung der lipomatösen Neoplasien, wurden, der Weltgesundheitsorganisation (WHO) folgend, bestimmte Kriterien als Unterscheidungsmerkmale angewendet. Die Klassifizierung als atypischer lipomatöser Tumor (ALT) beschreibt Tumoren, die histologisch komplett oder teilweise aus reifen adipozytären Proliferationen bestehen und Variationen der Zellgrößen enthalten.

Fokale nukleäre Atypien in Adipozyten und Stromazellen sowie MDM2-Amplifikationen sind hierbei vorhanden. Lipomatöse Neoplasien hingegen, die keine Amplifikation des MDM2-Gens aufweisen und aus reifen weißen Adipozyten bestehen, werden als Lipome klassifiziert (Jo & Fletcher, 2014). Auch die Lokalisation des Tumors spielt eine Rolle für dessen Einordnung. Während ALTs definitionsgemäß mit einem ausreichend großen Resektionsrand entnommen werden können, treten gut differenzierte Liposarkome (WDLs) retroperitoneal auf und/oder können nicht mit einem ausreichend großen Resektionsrand entnommen werden. In dieser Arbeit wurden keine WDLs eingeschlossen.

## 1.2 Einführung in die künstliche Intelligenz

### 1.2.1 Künstliche Intelligenz, maschinelles Lernen und Deep Learning

Ein Großteil der Forschungsarbeit im Bereich der künstlichen Intelligenz befasste sich in der Vergangenheit mit dem „Feature Engineering“, bei dem ganz konkrete

und von einem Experten festgelegte Merkmale berechnet werden. Dies führt wiederum zu Algorithmen, die auf einer Art Entscheidungsbaum basieren (LeCun et al., 2015).

Die künstliche Intelligenz ist dabei der Überbegriff, die den regelbasierten Ansatz beschreibt, bei dem versucht wird, Computer unter Zuhilfenahme einer Reihe von Regeln (wenn A dann B) für die Lösung von Problemen zu verwenden. Dieser Ansatz funktioniert nur bei sehr einfachen Problemen, beispielsweise in klar definierten Funktionen. Zu den dieser Art von KI zugrundeliegenden Informationen gehören unter anderem Expertenmeinungen und Publikationen (Hosny et al., 2018).

Ein Teilbereich der künstlichen Intelligenz (KI) ist das maschinelle Lernen. Hierunter finden sich Ansätze wie das „supervised learning“, das „unsupervised learning“ sowie das „reinforcement learning“. Bei dieser Forschungsarbeit wurde auf den Ansatz des „supervised learning“ gesetzt.

Beim maschinellen Lernen werden zur Lösung von Problemen neuronale Netzwerke verwendet. Diese Netzwerke haben eine variable Anzahl an Ebenen. Bei Verwendung vieler Ebenen (layer) wird die KI der nächsten Gruppe zugeordnet, wobei die Grenzen fließend verlaufen. Die hierbei zugrundeliegenden Informationen sind Daten (Hosny et al., 2018). Als Beispiel kann hier die Spracherkennung genannt werden, die heute in vielen Geräten Anwendung findet. Diese sind vorher mit großen Sprachdatensätzen trainiert worden.

Deep Learning ist eine maschinelle Lerntechnik, die von Experten festgelegte Merkmale vermeidet, indem sie die prädiktiven Merkmale aus Bildern erlernt und sich eigene Merkmale zur Unterscheidung sucht. Voraussetzung hierfür sind eine große Anzahl an Bilddatensätzen, die Annotationen von repräsentativen Merkmalen enthalten(Soffer et al., 2019). Durch Maschinelles Lernen können so ggf. Zusammenhänge entdeckt werden, die vorher nicht bekannt waren. Eine Visualisierung der Zusammenhänge ist in Abbildung 1 dargestellt.

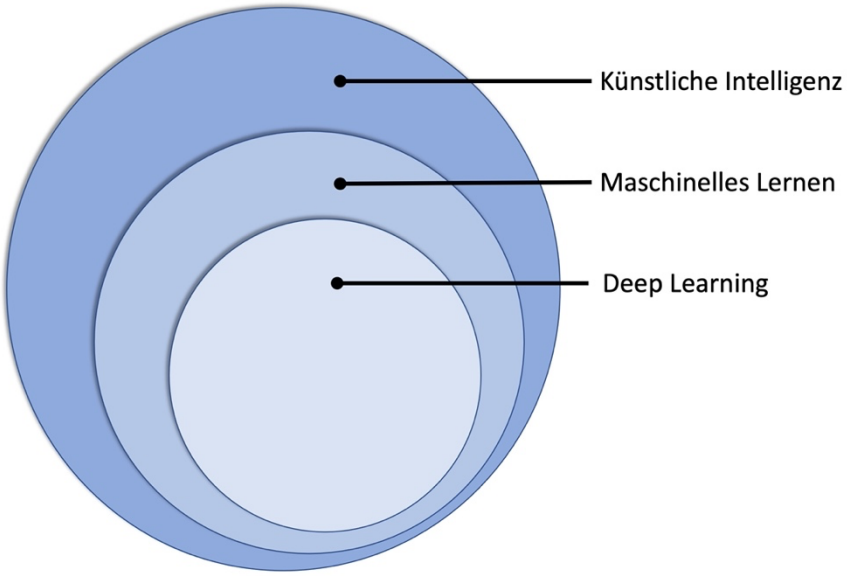


Abbildung 1: Visualisierung der Zusammenhänge von künstlicher Intelligenz

KI stellt einen Algorithmus dar, der Probleme selbst lösen kann. Maschinelles Lernen sind Algorithmen, die von Daten lernen können, ohne dafür speziell programmiert worden zu sein. Deep Learning ist ein Teilbereich des maschinellen Lernens, der auf neuronalen Netzen basiert.

### 1.2.2 Neuronale Netze

Maschinelles Lernen und insbesondere das Deep Learning basieren auf neuronalen Netzen. Das Ziel, das Wissenschaftler beim Bau der neuronalen Netze (NN) verfolgten, war, das menschliche Gehirn nachzubauen. Daher gibt es viele Gemeinsamkeiten zwischen artifiziellen und biologischen neuronalen Netzen wie dem Gehirn. Die artifiziellen NN bestehen in ihrer kleinsten Einheit aus dem künstlichen Neuron, auch Perzeptron genannt. Analog zur Nervenzelle gibt es auch hier mehrere Inputs und einen Output.

Beim NN wird zu Anfang ein Schwellenwert definiert, ab dem das künstliche Neuron ein Signal über seinen Output ausgibt, wie in Abbildung 2 und 3 dargestellt.

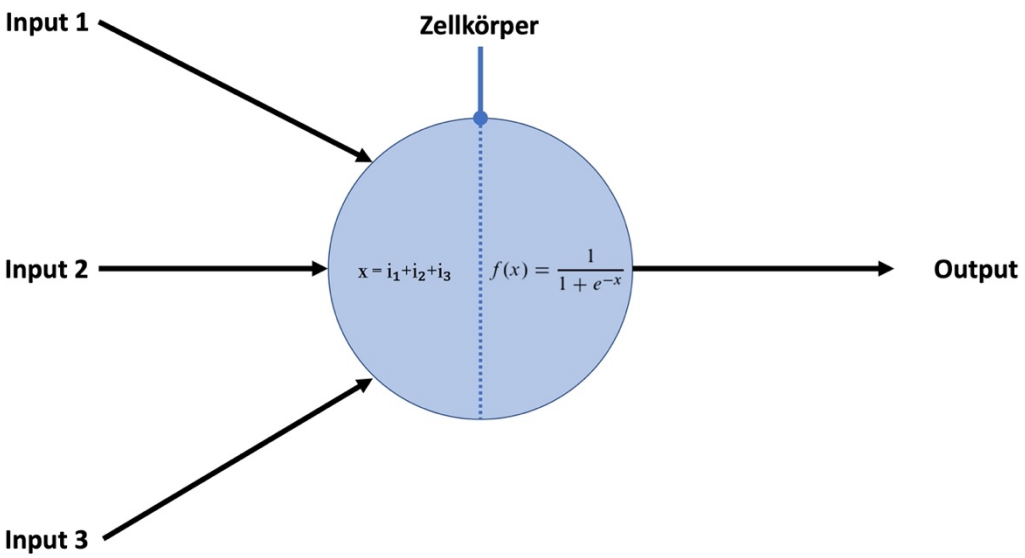


Abbildung 2: Schematische Darstellung eines künstlichen Neurons (Perzeptron) mit zugehöriger Schwellenwertfunktion (Sigmafunktion) zur Berechnung des Schwellenwertes. Die Funktion ist im Zellkörper auf der rechten Seite abgebildet. Alternative Schwellenwertfunktionen wären „tanh“ (-1 bis +1) und „ReLU“ (lineares Ansteigen des Outputs).

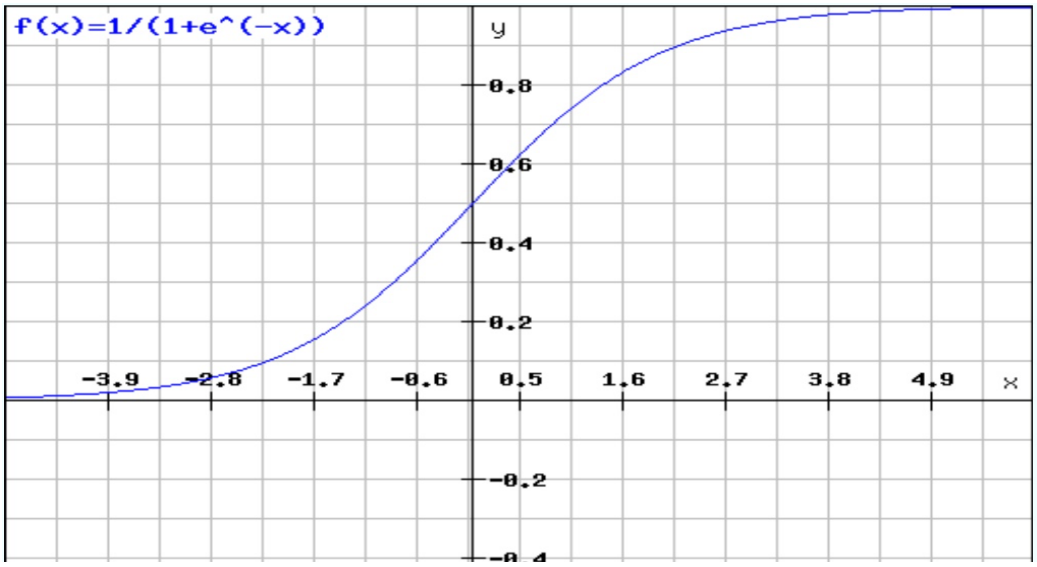


Abbildung 3 Sigmafunktion des Schwellenwertes beim künstlichen Neuron aus Abbildung 2 (grafische Darstellung).

Beim Menschen bestehen die Neuronen aus dem Zellkörper, dem Dendriten, über den Signale eingehen und einem Axon, über das Signale ausgegeben werden können.

Künstliche neuronale Netzwerke bilden im Verbund verschiedene Ebenen, die „layer“, wobei ein „input layer“, mehrere „hidden layer“ und ein „output layer“ unterschieden werden.

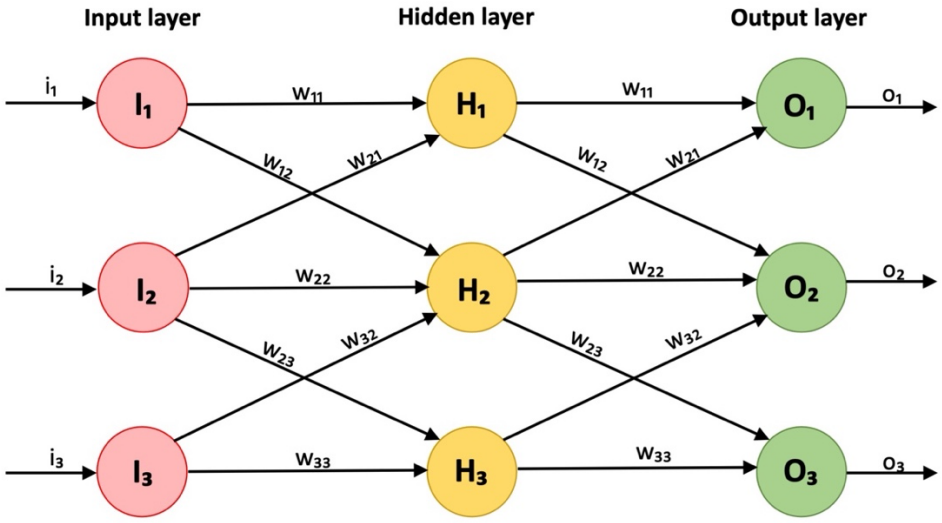


Abbildung 4: Beispiel eines neuronalen Netzes mit nur einem „hidden layer“. „i“ bezeichnet den Input, der in das NN eingespeist wird. „w“ bezeichnet die Gewichtungen, mit Werten zwischen 0 und 1, und somit die Stärke oder Wichtigkeit der neuronalen Verbindung. Der Buchstabe „o“ bezeichnet den Output und damit die Signale, die ausgegeben werden.

Im NN werden Informationen in Vektoren verarbeitet. Somit wird aus den drei Inputs des Beispiels i1, i2 und i3 ein Vektor, wie in Abbildung 4 dargestellt. Dieser wird nun mit einer Matrix multipliziert, mit der die Wichtungen „w“ eingehen. Die entsprechende Berechnung, ist in Abbildung 6 dargestellt. Diese Gewichtungen bestimmen den Output und damit das Ergebnis, das aus dem NN entsteht. Um dies zu veranschaulichen, wurde in Abbildung 5 ein NN mit Inputs und Wichtun-

gen versehen und der erste Schritt der Datenverarbeitung dargestellt. Zu beachten ist hierbei, dass, der Konvention folgend, die Schwellenwertfunktion (Sigmafunktion) noch nicht auf den „input layer“ angewendet wurde.

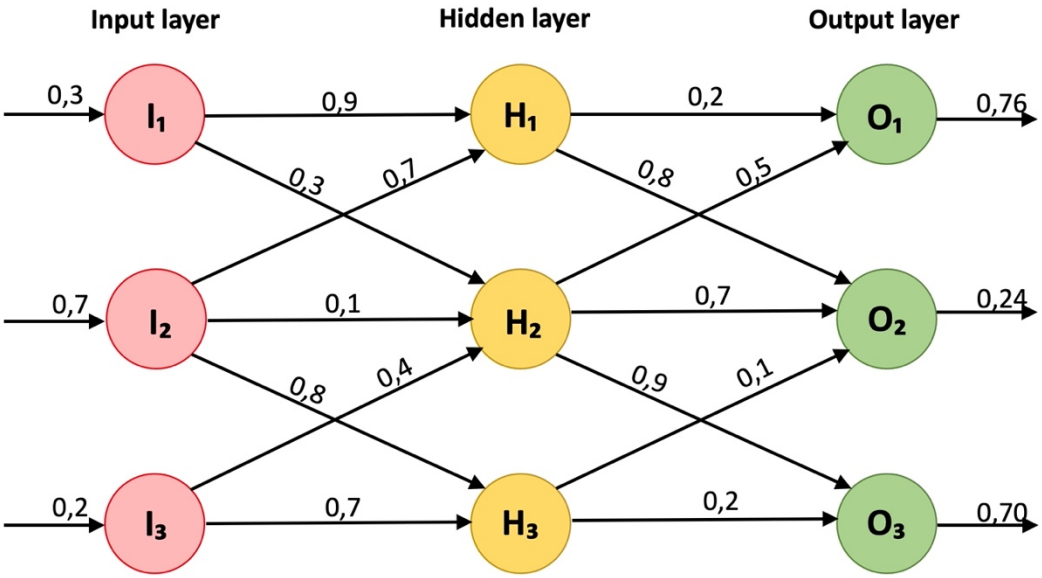


Abbildung 5 : Beispiel eines neuronalen Netzes mit nur einem „hidden layer“ und eingesetzten Gewichten.

Vektor Matrix Vektor

$$\begin{pmatrix} 0{,}3 \\ 0{,}7 \\ 0{,}2 \end{pmatrix} \cdot \begin{pmatrix} 0{,}9 & 0{,}7 & 0 \\ 0{,}3 & 0{,}1 & 0{,}4 \\ 0 & 0{,}8 & 0{,}7 \end{pmatrix} \approx \begin{pmatrix} 0{,}76 \\ 0{,}24 \\ 0{,}70 \end{pmatrix}$$

Abbildung 6 : Darstellung einer Beispielrechnung für das oben gezeigte neuronale Netz. Die Multiplikation eines Vektors mit einer Matrix ergibt wieder einen Vektor.

### 1.2.3 Die Gewichtung „w“ in neuronalen Netzen

Die Gewichtungen „w“, mit der die Stärke oder Wichtigkeit einer Verknüpfung ausgedrückt werden, sind die einzigen Stellschrauben in einem neuronalen Netz und damit der Teil, der angepasst wird, um die gewünschte Vorhersage zu bekommen.

### 1.2.4 Der Lernprozess des neuronalen Netzwerkes

Die bisher beschriebenen Abläufe beschreiben den „forward pass“ und somit den Weg von der Eingabe der Daten in den Input über die „hidden layer“ bis hin zum Output. Um ein NN aber lernfähig zu machen, wird eine Art Feedback benötigt, um erkennen zu können, wie weit das Ergebnis vom gewünschten Resultat entfernt ist. Daraufhin ist es dem NN möglich, die Gewichtung so anzupassen, dass eine Annäherung an das gewünschte Ergebnis („desired output“) möglich ist.

Wenn das NN das erste Mal auf ein Problem angewendet wird, kommt beim ersten „forward pass“, also beim erstmaligen Durchlaufen des NN vom Anfang („input“) bis zum Ende („output“), ein beliebiger Wert heraus. Dieser kann nun mit dem gewünschten Ergebnis („desired output“) verglichen werden. Um den Fehler möglichst gering werden zu lassen, wird eine Fehlerfunktion erstellt. Anschließend wird der Tiefpunkt dieser Funktion mithilfe des Gradientenabstiegsverfahren („gradient decent“) bestimmt, der zum Kern eines NN gehört. Die Formel hierfür, wird in Abbildung 7 dargestellt. Das Gradientenabstiegsverfahren verändert also die Gewichtung „w“ so, dass der Fehler minimiert wird. Im Gegensatz zum „forward pass“ verwendet das NN einen Optimierungsalgorithmus, um anzugeben, wie die Maschine die internen Gewichtungen ändern sollte, um die gewünschte Ausgabe eines Bildes am besten vorherzusagen. Der in dieser Technik verwendete Optimierungsalgorithmus nennt sich Backpropagation. Er wird verwendet, um anzugeben, wie eine Maschine ihre internen Parameter ändern sollte, um die gewünschte Ausgabe eines Bildes am besten vorherzusagen (LeCun et al., 2015).

$$W := W + \propto \cdot \nabla L$$

Abbildung 7: Das Gradientenabstiegsverfahren in einem neuronalen Netz

W Gewichte „w“ (W: =W ist ein Auftrag an den Computer, der besagt: Überschreibe das W vor dem „=“ mit W nach dem „=“)

α Lernrate (Schrittgröße)

$\nabla L$ Loss Gradient (beschreibt, wie die Gewichte anzupassen sind, damit der Fehler möglichst klein wird)

### 1.2.5 Bildverarbeitung in neuronalen Netzen

Für die NN werden Daten in Form von Vektoren benötigt. Deswegen müssen die Daten zur Bildverarbeitung, wie sie in dieser Forschungsarbeit verwendet wird, in einer bestimmten Form extrahiert werden. Hierzu wird das zu analysierende Bild in einzelne Pixel zerlegt, von denen jeder einzelne einen Input darstellt. Bei sehr großen Bildern werden Pixel zusammengelegt, um die Inputdatenmenge einzugrenzen.

### 1.2.6 Herkunft und Verbesserung der NN für den medizinischen Bereich

Die Deep-Learning-Algorithmen (DL), die in dieser Forschungsarbeit verwendet wurden, stammen von einem in ImageNet vortrainierten DenseNet-Algorithmus für das 2D-Modell und einem ResNet-Algorithmus für das 3D-Modell (Claudio E. von Schacky*, 2021; Huang et al., 2017). Die branchenübergreifende Anwendbarkeit und die dadurch bedingte häufige Nutzung dieser Technologie sorgen für eine schnelle Verbreitung und Verbesserung.

Die 3D-Algorithmen werden mit großen Investitionen und Anstrengungen vorangetrieben. Eine möglichst genaue Bildanalyse ist in vielen Bereichen äußerst wichtig. Wie im Bereich der medizinischen Bildanalyse, ist es bei der Entwicklung der Bilderkennung zum autonomen Fahren wichtig, keine falsch-negativen Ergebnisse zu liefern, da diese schwerwiegende Folgen wie das Übersehen einer Pathologie in einem Bilddatensatz haben könnten.

Ebenso wie der in dieser Forschungsarbeit verwendete Algorithmus auf einem DenseNet-Algorithmus und einem ResNet-Algorithmus basiert, werden kommende 3D-Bildanalysealgorithmen in der Medizin stark von der Weiterentwicklung in anderen Anwendungsbereichen profitieren.

Die Vorteile des Deep Learning bei der Bildanalyse in der Medizin wurden in zahlreichen Studien dargelegt. Es wurde erfolgreich auf den Bereich der muskuloskelettalen Radiologie sowie der Analyse von Mammografie-Aufnahmen angewendet (McKinney et al., 2020; Norman et al., 2018; von Schacky et al., 2020). Es wurde daraufhin die Hypothese aufgestellt, dass DL-Algorithmen in der Lage sein könnten, basierend auf Bildern zwischen Atypischen lipomatösen Tumoren (ALTs) und Lipomen zu unterscheiden. Bisher waren dafür eine histopathologische Untersuchung und eine immunhistochemische Analyse notwendig. Das Ziel dieser Forschungsarbeit ist die zuverlässige Unterscheidung von ALTs und Lipomen unter Zuhilfenahme von für diesen Anwendungsfall trainierten DL-Algorithmen in der Magnetresonanztomografie (MRT). Um die Richtigkeit der Diagnose sicherzustellen, wurde die Histopathologie und der „Mouse double minute 2“-Status (MDM2-Status) des Tumors bestimmt. MDM2 ist ein regulatorisch wirkendendes Gen, das bei hoher Amplifikation das Tumorsuppressor-Gen p53 hemmt. Bei Lipomen ist meist keine Überexpression des MDM2 zu beobachten, bei ALTs hingegen schon.
Um die Leistungsfähigkeit des DL-Modells einschätzen zu können, wurde anschließend die Leistung des Modells mit der Leistung von Assistenzärzten in der Radiologie, Fachärzten für Radiologie und Fachärzten mit Spezialisierung auf dem Gebiet der muskuloskelettalen Radiologie verglichen.

Besonders hervorzuheben ist dabei, dass bei der Bildanalyse der künstlichen Intelligenz andere systematische Analysefehler begangen wurden als bei der Analyse von Bildern, die von Radiologen ausgewertet wurden. Bei der Auswertung von Mammografie-Aufnahmen wurden beispielsweise einige wenige Fälle von Brustkrebs von ↑den teilnehmenden Radiologen übersehen, von der KI hingegen erkannt (McKinney et al., 2020). Umgekehrt kam es in derselben Studie vor, dass die KI einige Fälle übersehen hat, die für alle an der Studie teilnehmenden Ra-

diologen sehr einfach zu erkennen waren. Die Performanz der KI war in der vorangehenden Studie auf dem gleichen Niveau wie die Performanz der Radiologen.

Durch die Implementierung von KI in die von Radiologen vorgenommene Bildanalyse lassen sich somit synergistische Effekte erreichen.

### 1.2.7 Radiomics-Ansatz im Vergleich zu Deep Learning

Eine Radiomics-Analyse von Bildern basiert auf einer Pipeline, in der eine Vielzahl von manuell ausgewählten Merkmalen extrahiert werden. Anschließend findet eine Merkmalsselektion statt und eine auf maschinellem Lernen basierende Klassifizierung. Der Nachteil dieser Methode besteht darin, dass sich die Methode auf den aktuellen Wissensstand beschränkt, und somit Merkmale, die noch nicht bekannt, aber relevant sind, nicht genutzt werden. Beim Deep-Learning-Ansatz erfolgt diese Merkmalsauswahl nicht durch den Menschen, sondern die Kriterien werden vom Algorithmus selbst ausgewählt und können, wie oben erwähnt, mittels Gradient-weighted Class Activation Mapping visualisiert werden. Durch diese automatische Erkennung kann die Detektionsrate erhöht werden (Soffer et al., 2019; Sun et al., 2020). Es konnte gezeigt werden, dass Deep-Learning-Ansätze bei der Differenzierung zwischen gutartigen und bösartigen Läsionen der Brust Radiomics-Ansätzen überlegen sein können bei der Analyse von multiparametrischen MRT-Bildern (Truhn et al., 2019).

# 2. Zielsetzung

Das Ziel dieser Arbeit war die Entwicklung von 2D- und 3D-Deep-Learning-Ansätzen zur Differenzierung von Atypischen Lipomatösen Tumoren von Lipomen basierend auf magnetresonanztomografischen Bildern in verschiedenen Sequenzen. Die Ergebnisse sollen dann mit denen von Fachärzten und Assistenzärzten verglichen werden.

# 3. Material und Methoden

## 3.1 Patientenselektion

Vor der Studie wurde die Genehmigung der lokalen Ethikkommission (Ethikkommission der Technischen Universität München) eingeholt (Ethikantrag Nummer 666/21 S-KK). Es wurde eine retrospektive Analyse mit 260 Patienten durchgeführt, die die Klinik aufgrund einer lipomatösen Neoplasie der oberen Extremität, der unteren Extremität oder des Rumpfes aufgesucht haben. Für die Studie wurden alle Patienten mit entsprechendem Krankheitsbild, die zwischen Januar 2010 und Oktober 2018 im Klinikum rechts der Isar aufgenommen wurden, berücksichtigt. Bei allen Teilnehmern der Studie wurde präoperativ eine Bildgebung mittels MRT durchgeführt.

Bei 98 der Patienten wurde mittels histopathologischer Untersuchung der MDM2-Status nach Entnahme des Tumors unter Zuhilfenahme des Fluoreszenz-in-situ-Hybridisierung-Verfahrens untersucht (Knebel et al., 2019). Nur Patienten mit erhobenem MDM2-Status und komplettem MR-Datensatz wurden in die Studie aufgenommen. Eine entsprechende Darstellung ist in Tabelle 1 zu finden. Keiner der Patienten, die in die Studie aufgenommen wurden, hatte ein WDL.

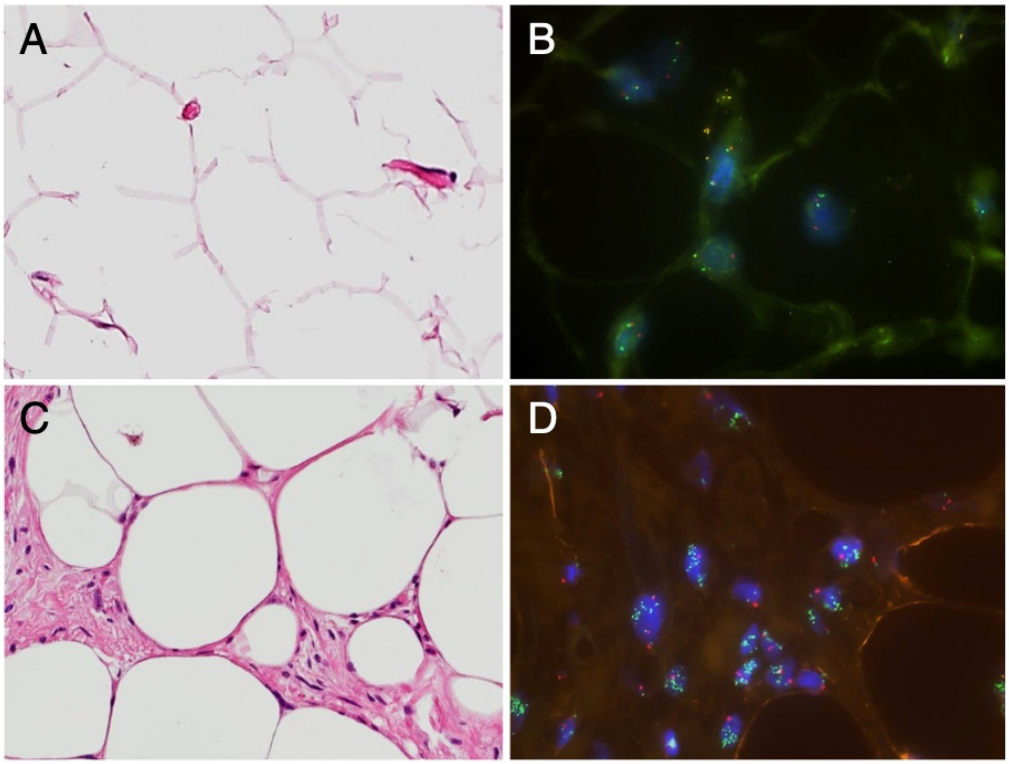


Abbildung 8: (a) Ein Lipom mit gleich großen Fettvakuolen, bei dem keine Atypien erkennbar sind. (b) Die Fluoreszenz-in-situ-Hybridisierungsanalyse (FISH) des gleichen Bereiches des MDM2-Gens (Disomie bezüglich MDM2, Grün: Gensonde MDM2-Region; Rot: Zentromersonde Chromosom 12; zwei grüne und zwei rote Signale pro Zelle bedeuten Disomie und somit keine Amplifikation des MDM2-Locus und somit präsentiert sich ein Lipom). (c) Atypische lipomatöse Tumoren (ALT) mit atypischen Stromazellen mit sichtbarer Kernhyperchromasie und unterschiedlich großen Fettvakuolen. (d) Entsprechende Fluoreszenz-in-situ-Hybridisierungsanalyse (FISH) des MDM2-Gens (Cluster-ähnliche Signale in Grün bedeuten eine Amplifikation des MDM2-Lokus, das rote Signal markiert das Zentromer von Chromosom 12 als Kontrolle. Hier präsentiert sich somit ein ALT (Knebel et al., 2019).

Die MRT wurde mit 1,5-Tesla- oder 3-Tesla-Scannern durchgeführt. Die untersuchten Sequenzen umfassten drei verschiedene Wichtungen. Eine T1-gewichtete Spin-Echo-Sequenz (T1w), eine T2-gewichtete Schnell-Spin-Echo-Sequenz (T2w) sowie eine T1w-Sequenz mit Fettunterdrückung nach der Gabe von Kontrastmitteln (T1fsgd). Die Bilddaten für diese Forschungsarbeit wurden in das Digital Imaging and Communications in Medicine Archiv als Neuroimaging-Informatics-Technology-Initiative (NIFTI)-Dateiformat für die Weiterverarbeitung aus dem 3D-Slicer extrahiert.

Die anschließenden Segmentierungen der Tumore wurden per Hand von D. W. K. unter der Anleitung von S. C. F. (einer Radiologin mit vierjähriger Erfahrung im Bereich der muskuloskelettalen Bildgebung) unter Verwendung der Open-Source-Software (3D-Slicer, Version 4.7; www.slicer.org) und von J. N. (einem Facharzt auf dem Gebiet der muskuloskelettalen Radiologie mit achtjähriger Erfahrung) durchgeführt. Eine Verblindung bezüglich Alter und Geschlecht wurde durchgeführt.

Die Segmentierungen wurden als NIFTI-Label-Maps für weitere Analysen in der zweidimensionalen (2D) und dreidimensionalen (3D) DL-Algorithmus-Pipeline aus dem 3D-Slicer extrahiert. Vor den Segmentierungen der 2D-Modelle wurde der Tumor zunächst betrachtet. Anschließend wurde manuell eine Schicht des Tumors ausgewählt, in der der Tumor einen möglichst großen Durchmesser hatte. Anschließend wurden im 3D-Slicer die Tumorränder exakt markiert und durch diese Umrandung die Abgrenzung des Tumors zum gesunden Gewebe definiert. In Abbildung 13 und Abbildung 14 ist dies beispielhaft dargestellt.

## 3.2 Entwicklung und Training des 2D- und 3D-Deep-Learning-Algorithmus

In dieser Forschungsarbeit wurden sowohl 2D- als auch 3D-DL-Algorithmen verwendet, um mithilfe der MRT-Bilddaten zwischen ALTs und Lipomen unterscheiden zu können. Der Einsatz bei der Klassifizierung medizinischer Bilddaten (Husseini et al., 2020; Jiménez-Sánchez et al., 2019; Navarro et al., 2018) als auch bei der Segmentierung (Navarro et al., 2019; Ronneberger et al., 2015) und der Lokalisation (Navarro et al., 2020; Xu et al., 2019) führte zu diesem Ansatz. Dadurch konnte einerseits die einzelne Leistung der DL-Algorithmen beurteilt und andererseits die Vergleichbarkeit untereinander ermöglicht werden.

### 3.2.1 2D-Deep-Learning-Modell

In der Forschungsarbeit wurde ein 2D-Deep-Learning-Modell verwendet, das auf einem in ImageNET vortrainierten Dense Convolutional Network (DenseNET-201) basiert (Huang et al., 2017). ImageNET ist ein großer Bilddatensatz, der unter Zuhilfe-nahme der WordNet-Hierarchie organisiert ist. Jedes Bild ist

durch Wörter oder Wortphrasen nominiert, die unter Zuhilfenahme menschlicher Qualitätskontrolle entstanden sind.

DenseNETs sind neuronale Netzwerke, die durch die Netzwerkarchitektur einige Vorteile bieten. Sie vermindern das Problem des verschwindenden Gradienten, verstärken die Merkmalsausbreitung, fördern die Wiederverwendung von Merkmalen und sorgen für eine Reduktion der Anzahl von Parametern (Deng et al., 2009; Huang et al., 2017). Der Workflow dieses Algorithmus wird in Abbildung 9 näher dargestellt.

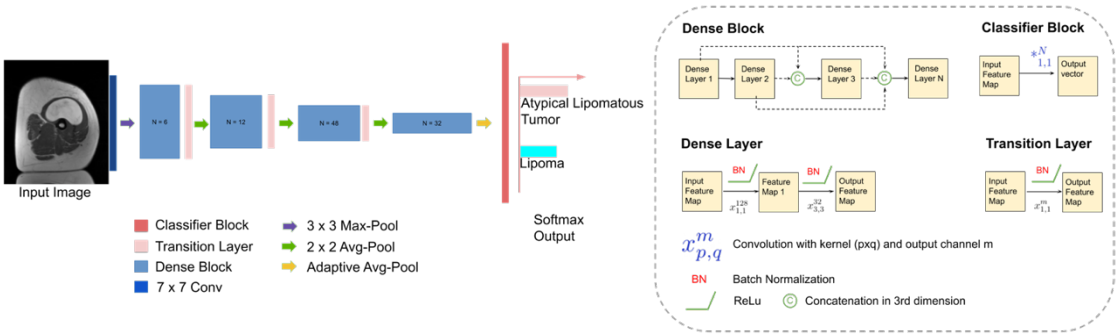


Abbildung 9: Architektur des DenseNet 201 (von Schacky et al., 2023; Status: eingereicht) 2D-Modell der DenseNet-201-Architektur zur Unterscheidung zwischen atypischen lipomatösen Tumoren und Lipomen. Das Netzwerk empfängt die ausgewählte 2D-Scheibe von jedem Scan und gibt die Wahrscheinlichkeit des Bildes für die beiden Klassen an. Auf der linken Seite werden die einzelnen Komponenten und Schichten des DenseNet beschrieben. Für weitere Einzelheiten siehe folgende Quellen: (Deng et al., 2009; Huang et al., 2017).

### 3.2.2 3D-Deep-Learning-Modell

Für die weitere Untersuchung, insbesondere ob die 3D-Kontextinformationen relevant für Differenzierung von Lipomen und ALTs sind, wurde ein 3D-Modell auf der Grundlage einer 2D-ResNet-Architektur erstellt (He et al., 2016). Eine bildliche Darstellung dieser ResNet-Architektur ist Abbildung 10 zu entnehmen.

ResNet charakterisiert ein residuales neuronales Netz, das einen DL-Algorithmus darstellt, der ähnlich wie die Pyramidenzellen des zerebralen Cortex aufgebaut ist. Die Besonderheit besteht darin, dass einzelne Schichten im Netzwerk übersprungen werden. Dies geschieht im Cortex bei Neuronen der kortikalen Schicht VI, die einen Input von Schicht I erhalten, unter Umgehung der Zwischenschichten (Thomson, 2010).

In dieser Forschungsarbeit wurde ein vierschichtiges dreidimensionales Netzwerk mit Restblöcken angewendet. Diese ermöglichen die Übertragung von Informationen aus verschiedenen vorausgehenden Schichten und nicht lediglich aus einer unmittelbar vorangegangenen Schicht. Dadurch sind tiefere Architekturen möglich, was zu genaueren Netzen führt.

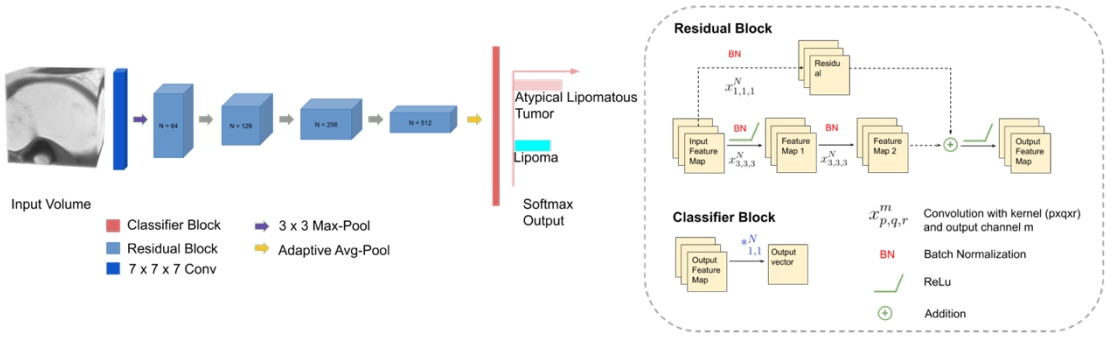


Abbildung 10: Architektur des ResNet 3D-ResNet-Architektur zur Unterscheidung zwischen atypischen lipomatösen Tumoren und Lipomen. Die Eingabe für das Netz ist eine 100×100x100 ROI, und es gibt die Wahrscheinlichkeit des Bildes für die beiden Klassen aus. Ganz links in der Abbildung ist ein Restblock beschrieben. Weitere Informationen sind unter folgender Quelle zu finden (He et al., 2016).

### 3.2.3 Weiterverarbeitung der Daten

Alle in dieser Arbeit verwendeten Volumina wurden zu einer isotropischen Auflösung von 0,5 $mm^3$ und mithilfe einer Z-Score-Normalisierung prozessiert. Für das 2D-Modell wurde die Schicht verwendet, die die größte Tumorfläche aufwies. Es stellte sich heraus, dass die Nutzung dieser Schicht den Tumor für das 2D-Modell am besten repräsentiert. Es wurden außerdem Anpassungen der Bildgröße vorgenommen, um die Anforderungen des von uns angewendeten vortrainierten NN zu erfüllen. Die Bildgröße wurde auf 224 x 224 Pixel angepasst. Für das 3D-Deep-Learning-Model wurden die Bilder auf die Größe von 100 x 100 x 100 Voxel angepasst. Es wurde außerdem eine Segmentierungsmaske verwendet, um die Region of Interest (ROI) auch nach der Beschneidung im Bild zu behalten (von Schacky et al., 2023; Status: eingereicht).

### 3.2.4 Training des Deep-Learning-Modells

Alle Modelle wurden mithilfe einer 16-GB-Tesla-V100-Grafikkarte in Pytorch entwickelt (Paszke et al., 2017). Die 2D-Modelle wurden mit einer Stapelgröße von 16 und einer Lernrate von $2^{-4}$ mit einem ADAMW-Optimierer trainiert. Die 3D-Modelle wurden mit einer Stapelgröße von 8 und einer Lernrate von $4^{-4}$ trainiert. Das Training des Algorithmus wurde mit frühzeitigem Abbruch und der Überwachung des Validierungsverlustes durchgeführt, um das beste Modell auszuwählen. Die angewandte Verlustfunktion stellte die kategorial gewichtete Kreuzentropie dar.
Die Datenerweiterung wurde zur Trainingszeit angewendet. Sie umfasste Spiegelungen (vertikal und horizontal), elastische Umformung, zufälliges Zoomen, zufällige Rotation, zufälliges gaußsches Rauschen, zufälliges Beschneiden, zufälliges Ghosting und zufälliges „Bias Field“ (Subramanian, 2018). Es wurden des Weiteren „Minority Class Oversampling“ während des Trainings angewendet (Selvaraju et al., 2017).
Weitere Details bezüglich des angewendeten Trainings für das NN sind online hinterlegt (https://github.com/ALTLipomaNet) (von Schacky et al., 2023; Status: eingereicht).

### 3.2.5 Evaluation des Deep-Learning-Modells

Wie bei der Entwicklung von NN üblich, wurden für die unabhängige Prüfung vor dem Training 20 % (n = 21) der Bilder entfernt, mithilfe derer der entstandene DL-Algorithmus getestet werden konnte. Mit den verbleibenden 80 % (n = 77) des Datensatzes wurde der DL-Algorithmus unter Anwendung einer 3-fachen Kreuzvalidierung trainiert und validiert.

Damit Stichprobenverzerrungen minimiert werden konnten, wurden die Kreuzvalidierungen dreimal wiederholt, was zu neun verschiedenen Modellen mit einer Hyperparameterkonfiguration führte. Die Modelle, die der Konfiguration mit der besten durchschnittlichen Leistung über alle Foldings (Faltungen) entsprachen, wurden für die finale Bewertung ausgewählt.

## 3.3 Bewertung der Auswertung der Testbilder durch Radiologen

Damit die Performanz der entwickelten DL-Modelle mit der Performanz der Radiolog*innen verglichen werden konnte, wurden Radiolog*innen mit unterschiedlichem Fachwissen hinsichtlich des Namens, Alters und Geschlechtes verblindet und es wurde von den Radiologen eine Einteilung der lipomatösen Neoplasien in Lipom oder ALT vorgenommen.
Die Bewertungen wurden von zwei Assistenzärzt*innen der Radiologie (beide im dritten Jahr), zwei Fachärzt*innen der Radiologie (beide in der Ausbildung zum Spezialisten für muskuloskelettale Radiologie) und vier Oberärzt*innen mit Spezialisierung auf muskuloskelettale Radiologie (A.S.G., J.N., A.A.F., K.W., mit 6, 6, 7 und 20 Jahren Erfahrung in der Bildgebung von Weichteiltumoren) durchgeführt.

## 3.4 Modellinterpretation und Datenvisualisierung

Um ein Bild davon zu bekommen, wie das DL-Modell arbeitet, wurden Heatmaps aus Gradienten-gewichteten Class Activation Maps (Grad-CAM) erstellt. Wie zuvor erwähnt, wurden mit diesen Heatmaps die Bereiche dargestellt, auf die sich der DL-Algorithmus bei seiner Entscheidung besonders konzentriert (Selvaraju et al., 2017).
Anhand von zwei Testreihen wurde bei den Heatmaps und den zugehörigen MR-Bildern überprüft, ob sich der DL-Algorithmus auf den Bereich des Tumors konzentriert. Hiermit konnten mögliche Ursachen für richtige und fehlerhafte Bewertungen, wie beispielsweise das „Overfitting“, analysiert werden.

## 3.5 Statistische Auswertung

Im nächsten Schritt wurden die verschiedenen Modelle bewertet. Hierzu wurde die Area under the curve (AUC) aus der Receiver-Operating-Characteristic-Analyse (Grenzwertoptimierungskurve oder ROC-Analyse) jeweils für den Validierungssatz und den unabhängigen Hold-out-Testsatz ermittelt.

Wie zuvor beschrieben, wurde die endgültige Leistung des Modells anschließend anhand der zuvor entnommenen 20 % der Bilddaten auf Genauigkeit, Sensitivität, Spezifität und AUC bewertet (Powers, 2020).

Die Berechnungen der Modellmetriken und Modellbewertungen, wurden mit scikit-learn, statsmodels 0,9 (statsmodels.org, Open Source) und dem pROC-Paket in R 3.6 mit der Methode nach DeLong durchgeführt (Robin et al., 2011).

Die Konfidenzintervalle wurden dann mithilfe des Perzentil-Bootstrap mit k = 1000 Iterationen berechnet.

# 4. Ergebnisse

## 4.1 Patientendaten und Datensätze

Für die Studie wurden 98 Patienten mit einem Durchschnittsalter von 53 ± 13,1 Jahren eingeschlossen. Davon waren 56 Patienten Frauen (57,1 %). Eine Übersicht davon ist auf Tabelle 1 zu sehen. Unter Verwendung des MDM2-Status als Referenzstandard, wurde bei 55 Patienten (56 %) ein Lipom und bei 43 (44 %) ein ALT diagnostiziert. Für die Durchführung von Training und Validierung mit 3-facher Kreuzvalidierung wurden 77 Patienten der Gesamtpopulation zufällig ausgewählt (Durchschnittsalter 54,9 ± 14,9 Jahre, 43 Frauen). 21 Patienten bildeten die unabhängige Hold-out-Testgruppe für den Vergleich durch Radiolog*innen (Durchschnittsalter 54,3 ± 12,3 Jahre, 15 Frauen). Hierrunter befanden sich 9 ALTs und 12 Lipome (von Schacky et al., 2023; Status: eingereicht). Ein Überblick über die Leistung der 2D-Deep-Learning-Modelle gibt die Abbildung 12.

**Tabelle 1:** Patientenmerkmale*

| **Patientenmerkmale** | **Gesamt** (n = 98) | **Training/Validierungssatz Für „3-fold Cross Validation“** (n = 77) | **Unabhängiger Testsatz** (n = 21) |
|---|---|---|---|
| **Alter*** (Jahre) | 53 (1,4) | 53 (1,6) | 52 (2,8) |
| **Geschlecht** (Frauen) | 56 | 43 | 13 |
| **Ort** | | | |
| Obere Extremität | 15 | 11 | 4 |
| Untere Extremität | 83 | 66 | 17 |
| **Lipome** | n = 55 | n = 43 | n = 12 |
| Alter (Jahre) | 55 (1,7) | 53 (2,0) | 51(3,3) |
| Geschlecht (Frauen) | 27 | 20 | 7 |
| **Atypische lipomatöse Tumoren** | n = 43 | n = 34 | n = 9 |
| Alter (Jahre) | 58 (2,2) ± 12,6 | 58 (2,5) | 53 (4,9) |
| Geschlecht (Frauen) | 29 | 24 | 5 |

Tabelle 1: Patientencharakteristika

*Daten als Median angegeben (Standardfehler).

Quelle: (von Schacky et al., 2023; Status: eingereicht)

## 4.2 Ergebnisse des 2D-Deep-Learning-Modells

Zunächst wurde die Performanz des entwickelten 2D-DL-Modells untersucht. Auf dem unabhängigen Testsatz ergab sich eine Genauigkeit von 95 % (95%-KI: 86 %, 100 %) bei 89 % Sensitivität (95%-KI: 62 %, 100 %) und 100 % Spezifität (95%-KI: 100%, 100 %) mit einer AUC von 0,98 (95%-KI: 0,91, 1,0) auf T1w-Bildern, wie in Abbildung 11 und Abbildung 12 dargestellt. Bei T2w-Bildern erreichte das 2D-DL-Modell eine Genauigkeit von 79 % (95%-KI: 57 %, 100 %) bei einer Sensitivität von 83 % (95%-KI: 44 %, 100 %) und einer Spezifität von 75 % (95%-KI: 43 %, 100 %) mit einer AUC von 0,85 (95%-KI: 58 %, 100 %). Bei T1fsgd-Bildern zeigte das 2D-DL-Modell eine Genauigkeit von 85 % (95%-KI: 62 %, 100 %) bei einer Sensitivität von 100 % (95%-KI: 100 %, 100 %) und einer Spezifität von 75 % (95%-KI: 44 %, 100 %) mit einer AUC von 0,82 (95%-KI: 0,5, 1,0). Tabelle 2 zeigt einen Überblick über die Leistung der 2D-Modelle auf T1w-Bildern der Testgruppe (von Schacky et al., 2023; Status: eingereicht).

## 4.3 Ergebnisse des 3D-Deep-Learning-Modells

Anschließend wurde die Leistung des 3D-DL-Modells betrachtet. Auf dem unabhängigen Testsatz ergab sich eine Genauigkeit von 82 % (95%-KI: 64 %, 95 %) bei 78 % Sensitivität (95%-KI: 44 %, 100 %) und 85 % Spezifität (95%-KI: 62 %, 100 %) mit einer AUC von 0,91 (95%-KI: 0,77, 1,0) auf T1w-Bildern, wie in Abbildung 12 dargestellt. Bei T2w-Bildern zeigte das 3D-DL-Modell eine Genauigkeit von 73 % (95%-KI: 47 %, 93 %) bei 67 % Sensitivität (95%-KI: 20 %, 100 %) und 78 % Spezifität (95%-KI: 50 %, 100 %) mit einer AUC von 0,69 (95%-KI: 0,37, 1,0). Bei T1fsgd-Bildern zeigte das 3D-DL-Modell eine Genauigkeit von 57 % (95%-KI: 29 %, 86 %) bei 80 % Sensitivität (95%-KI: 33 %, 100 %) und 44 % Spezifität (95%-KI: 12 %, 78 %) mit einer AUC von 0,64 (95%-KI: 0,34, 0,95). Tabelle 2 und Abbildung 12 geben einen Überblick über die erreichte Leistung der 3D-Modelle auf T1w-Bildern in der Testgruppe.

Abbildung 12 zeigt die ROC-Kurven für alle entwickelten Modelle. In der Forschungsarbeit fiel auf, dass die AUCs für 2D-Ansätze höher als für 3D-Ansätze. Der hier verwendete 2D-Ansatz mit T1w-Bildern als Input ergab mit 0,97 die höchste AUC (von Schacky et al., 2023; Status: eingereicht).

## 4.4 Vergleich mit den Erkennungsraten von Radiologen

Zwei Assistenzärzt*innen für Radiologie erreichten eine durchschnittliche Genauigkeit von 62 % bei 43 % Sensitivität und 75 % Spezifität. Wie erwartet, übertrafen die vier Fachärzt*innen für Radiologie die weniger erfahrenen Radiolog*innen. Diese erreichten eine durchschnittliche Genauigkeit von 74 % bei einer Sensitivität von 76 % und einer Spezifität von 69 %. Die Einzelergebnisse der Fachärzt*innen für Radiologie waren für Genauigkeit/Sensitivität/Spezifität 74%/76%/69%, 74%/65%/75%, 94%/100%/93% und 94%/87%/100%. Im Vergleich dazu übertraf das beste DL-Modell die Leistung von Assistenzärzt*innen, brachte aber etwa die gleiche Leistung wie Fachärzt*innen für Radiologie (von Schacky et al., 2023; Status: eingereicht). Ein Vergleich der Erkennungsraten ist in Tabelle 2 aufgeführt.

**Tabelle 2:** Leistung der Deep-Learning-Modelle und der Vergleich mit den Ergebnissen von Radiologen*innen*

| | **Genauigkeit** (%) | **Sensitivität** (%) | **Spezifität** (%) | **AUC** |
|---|---|---|---|---|
| **2D-Deep-Learning-Modell** | | | | |
| T1w | 95 | 89 | 100 | 0.98 |
| T2w | 79 | 83 | 75 | 0.85 |
| T1fsgd | 85 | 100 | 75 | 0.82 |
| **3D-Deep-Learning-Modell** | | | | |
| T1w | 82 | 78 | 85 | 0.91 |
| T2w | 73 | 67 | 78 | 0.69 |
| T1fsgd | 57 | 80 | 44 | 0.64 |
| **Auswertung Radiolog*innen** | | | | |
| Radiologie Assistenzärzt*in 1 | 62 | 43 | 75 | |
| Radiologie Assistenärzt*in 2 | 63 | 77 | 55 | |
| Radiologie Fachärzt*in 1 | 74 | 76 | 69 | |
| Radiologie Fachärt*in 2 | 74 | 65 | 75 | |
| Radiologie Fachärzt*in 3 | 94 | 100 | 93 | |
| Radiologie Fachärzt*in 4 | 94 | 87 | 100 | |

Tabelle 2: Leistung des entwickelten zweidimensionalen (2D) und dreidimensionalen (3D) Deep-Learning-Modells, gemessen mit Genauigkeit, Sensitivität, Spezifität und Area under the curve (AUC). Die Leistung wird in Abhängigkeit von der verwendeten Sequenz angegeben: T1-gewichtete Spin-Echo-Sequenz (T1w), T2-Schnell-Spin-EchoSequenz (T2w), Kurztau-Inversionssequenz mit Fettunterdrückung nach Anwendung eines Kontrastmittels auf Gadolinium-Basis (T1fsgd). Zum Vergleich mit den durch Radiologen erzielten Ergebnissen von wurden die MRT von zwei Assistenzärzt*innen für Radiologie und vier Fachärzt*innen für Radiologie, die hinsichtlich histologischer und klinischer Daten verblindet waren. Die Bewertung stützt sich auf Merkmale der MR-Bildgebung wie dicke Septen, Tumordurchmesser, dem Vorhandensein von Knoten und Kontrastverstärkung.
Quelle: (von Schacky et al., 2023; Status: eingereicht)

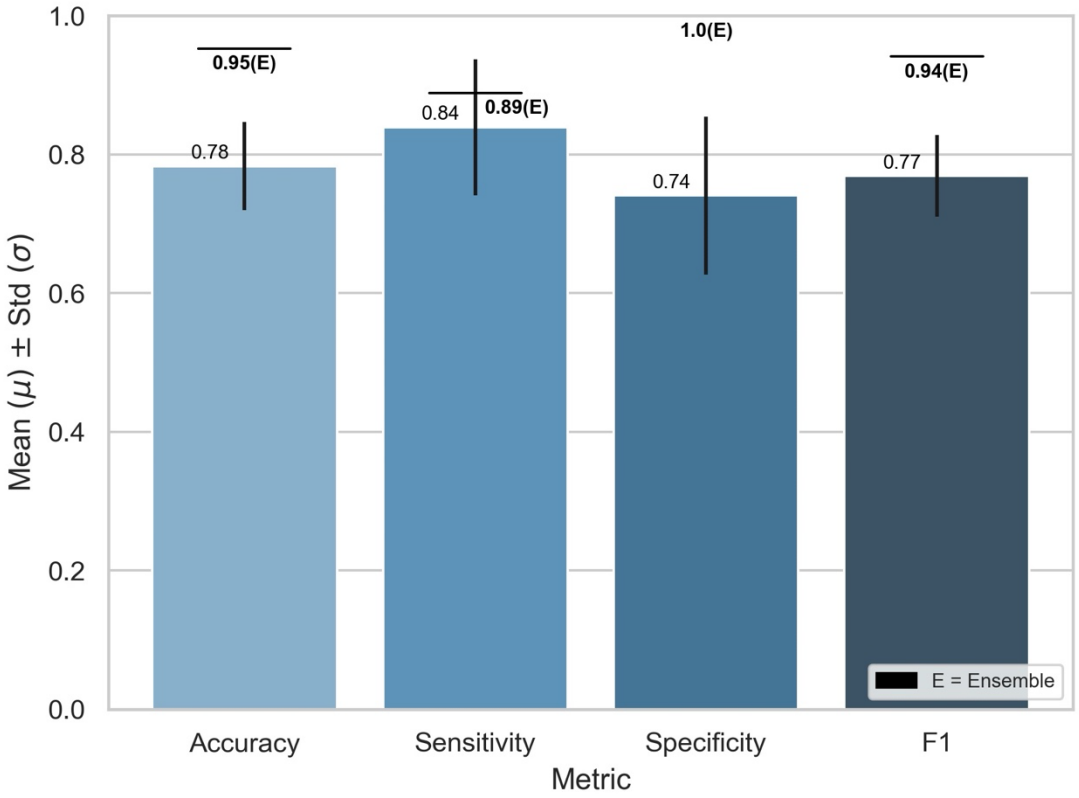


Abbildung 11: Leistung des zweidimensionalen (2D) Modells auf den T1w-Bildern im Testsatz: Genauigkeit, Sensitivität und Spezifität, sowie der F1-score jedes Modells (Median und Satndardabweichung). Auch die Kombination der Modelle wurde dargestellt und mit dem Buchstaben E markiert auf dem Diagramm dargestellt. Diese Kombination führte zu einer Genauigkeit von 95%, einer Sensitivität von 89% und einer Spezifität von 100%, sowie einem F1-score von 94.

Quelle: (von Schacky et al., 2023; Status: eingereicht)

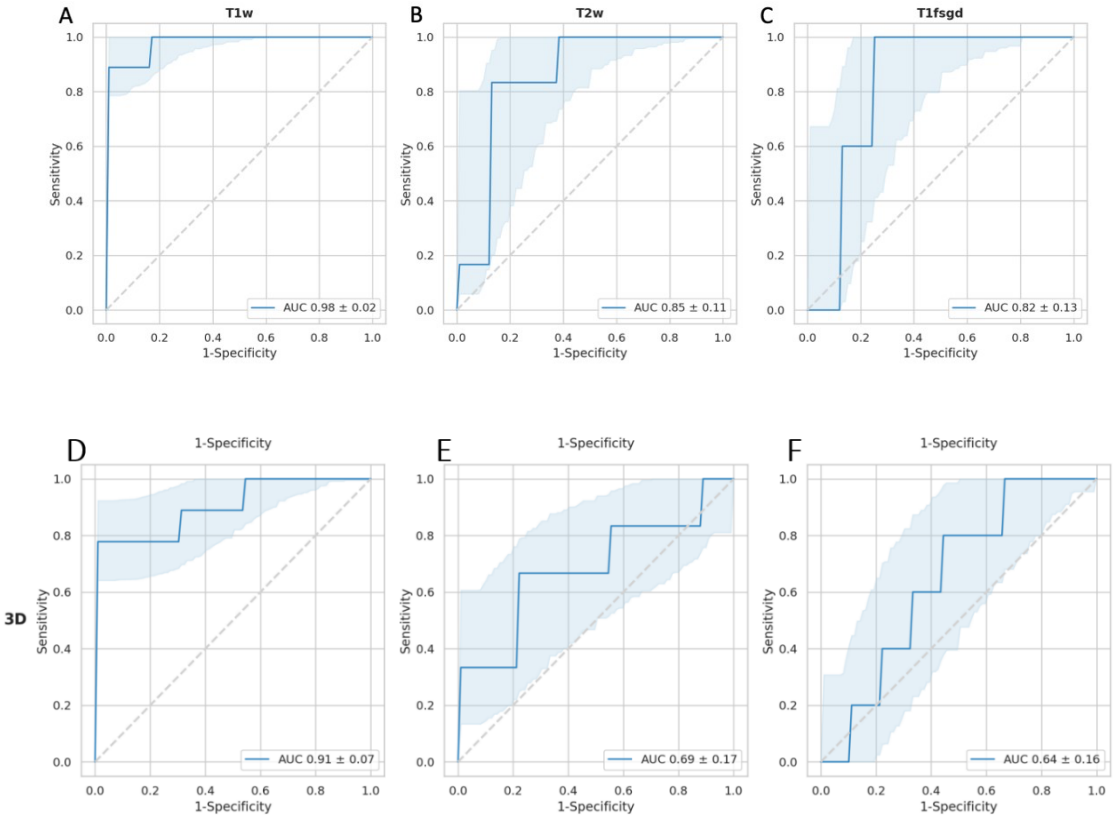


Abbildung 12: Receiver-Operating-Characteristics(ROC)-Kurven und die entsprechenden Area under the curve (AUC) von zweidimensionalen (2D) und dreidimensionalen (3D) Modellen auf dem Testsatz mit Standardabweichung:
A: ROC-Kurve des 2D-Modells mit T1-gewichteten Spin-Echo-MR-Bildern (T1w) als Input.
B: ROC-Kurve des 2D-Modells unter Verwendung von schnellen T2-Spin-Echo-MR-Bildern (T2w) als Input.
C: ROC-Kurve des 2D-Modells unter Verwendung von T1-gewichteten MR-Bildern mit Fettunterdrückung und nach Kontrastmittelapplikation (T1fsgd) als Input.
D: ROC-Kurve des 3D-Modells unter Verwendung von T1w als Input.
E: ROC-Kurve des 3D-Modells unter Verwendung eines T2w als Input.
F: ROC-Kurve des 3D-Modells unter Verwendung von T1fsgd als Input.
Quelle: (von Schacky et al., 2023; Status: eingereicht)

## 4.5 Interpretation von Modellen und Visualisierung

Nach der letzten Faltungsschicht wurden die Daten für die Grad-CAMs gewonnen, die danach mit den zugehörigen MR-Bildern überlagert wurden.
Somit konnte die Entscheidungsfindung des Algorithmus visualisiert werden. Wie in Abbildung 15 zu erkennen ist, konzentrierte sich der DL-Algorithmus überwiegend auf den Bereich des Tumors, genauer auf den Rand des Tumors. Dies kann durch die rote Färbung, die eine starke Aktivierung bedeutet, erkannt werden. Bei

den kontrastmittelverstärkten Bildern (T1fsgd) fand eine starke Aktivierung vor allem dort statt, wo eine starke Aufnahme an Kontrastmitteln erfolgte.
Eine inkorrekte Klassifizierung auf den mit der Grad-CAM überlagerten MR-Bildern wird in Abbildung 16 deutlich. Der DL-Algorithmus konzentrierte sich hier hauptsächlich auf Bereiche außerhalb des Tumors, aufgrund dessen eine inkorrekte Klassifikation erfolgte. Es ist zudem ersichtlich, dass sich der DL-Algorithmus zum Teil auf Bereiche außerhalb des Tumors konzentrierte, die Septen enthielten und dadurch zu einer inkorrekten Einordnung eines Lipoms als ALT geführt hat (von Schacky et al., 2023; Status: eingereicht)

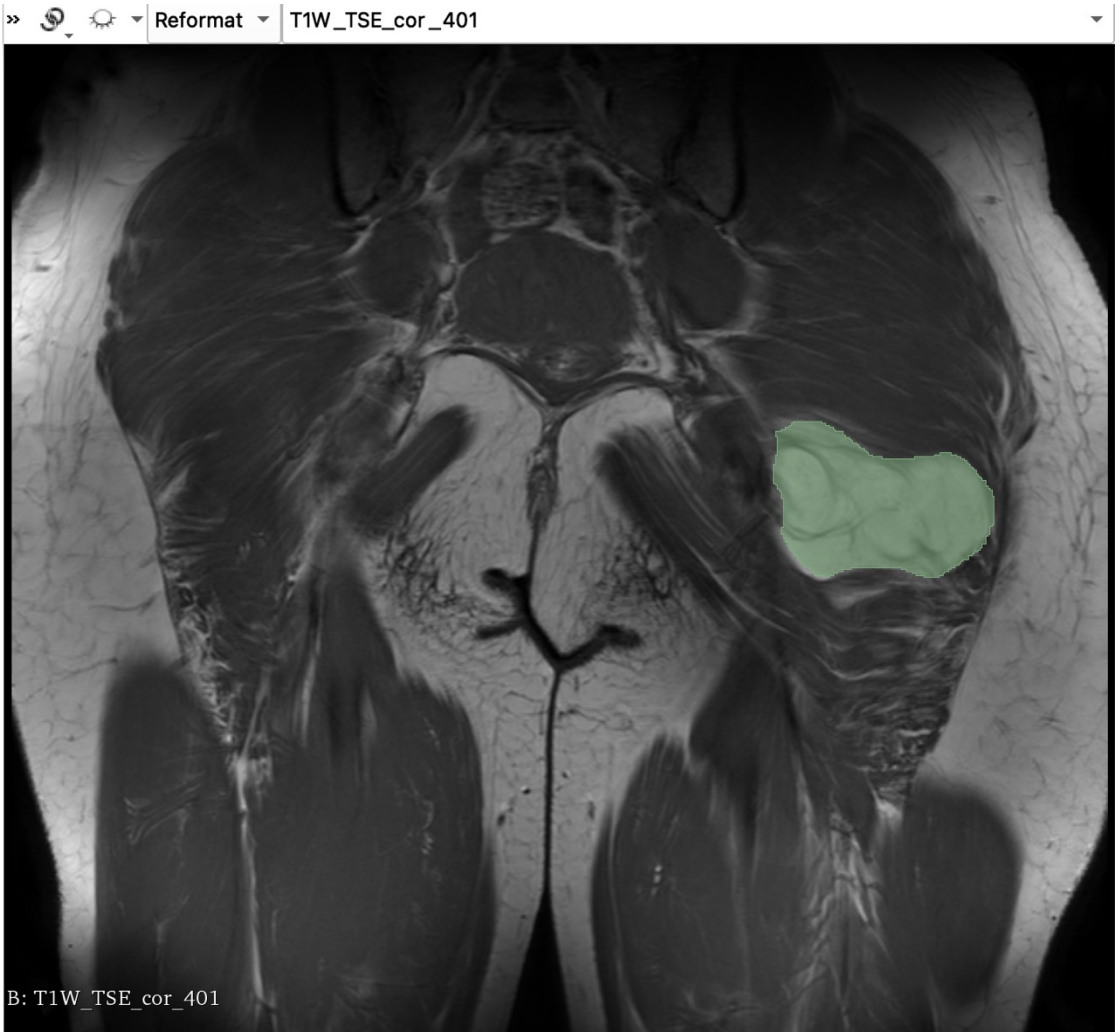


Abbildung 13: Darstellung der Oberfläche des Open-Source-Programmes 3D-Slicer, mit dem die Segmentierung der Tumoren durchgeführt wurde. Dieses Bild zeigt eine erste grobe Markierung des Tumors in T1-Wichtung, die mit verschiedenen Korrekturwerkzeugen anschließend weiter angepasst wurde.

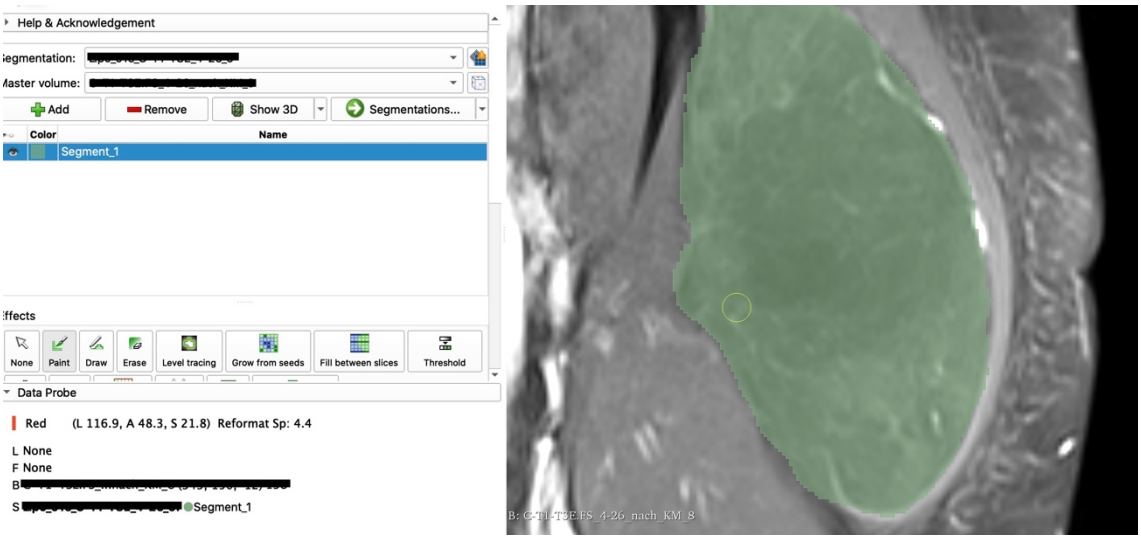


Abbildung 14: Darstellung der Oberfläche des Open-Source-Programmes 3D-Slicer, mit dem die Segmentierung der Tumoren durchgeführt wurde. Dieses Bild zeigt eine fertig segmentierte Schicht eines Tumors in T1w.

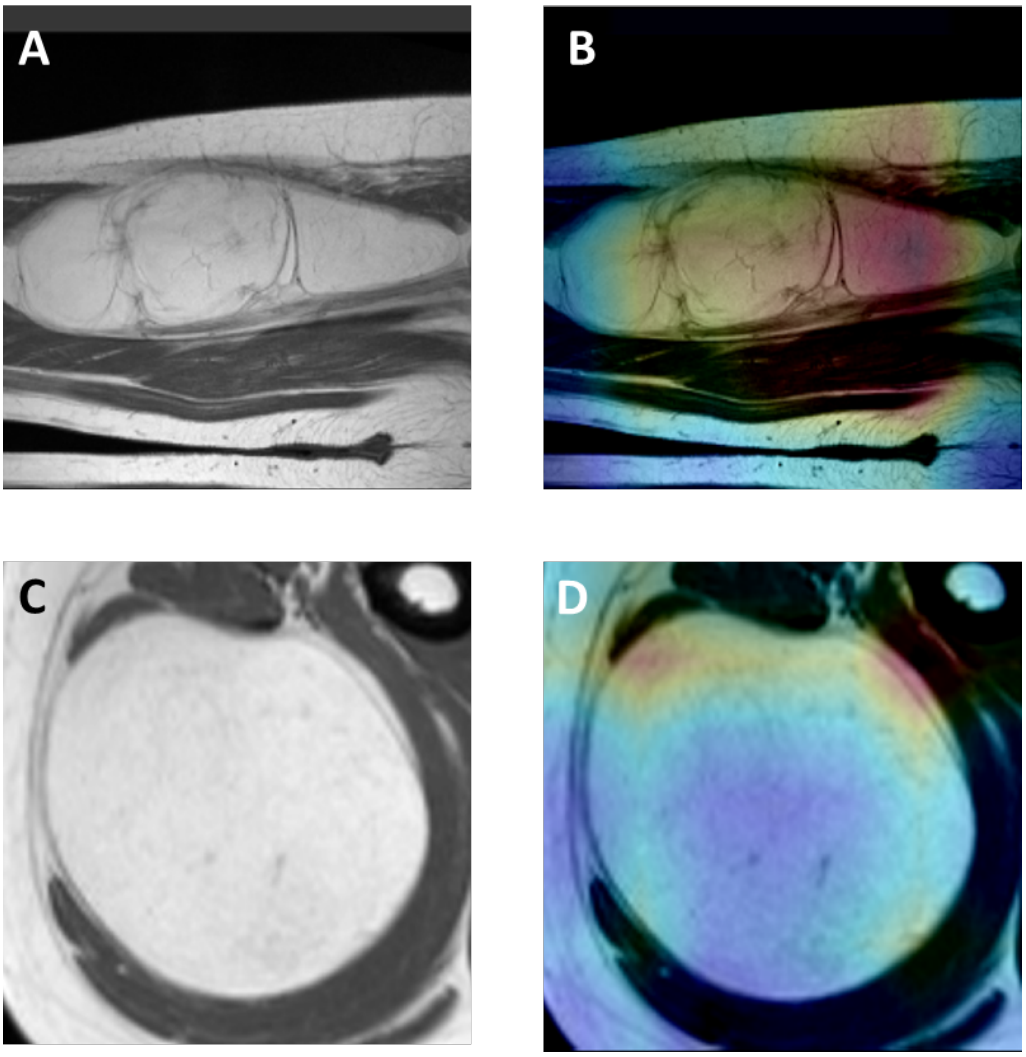


Abbildung 15: Hier sind zwei Beispiele dargestellt, in denen der DL-Algorithmus richtig klassifiziert hat:
Eine rote Färbung stellt eine hohe, eine blaue eine niedrige Aktivierung dar. Bei beiden Bildern handelt es sich um T1-gewichtete MR-Bilder. Die Bilder B und D stellen jeweils die von einer Grad-CAM überlagerten Input-Bilder des vorhergehenden Bildes dar.
Bei Bild A und Bild B ist ein ALT zu sehen, der mit positivem MDM2-Status verifiziert wurde. Das am besten funktionierende 2D-DL-Modell konnte diesen Tumor mit einer Sicherheit von 82 % als ALT klassifizieren. Gut zu erkennen sind die dicken Septen (> 2 mm), die innerhalb des Tumors verlaufen und häufig bei ALTs zu finden sind. Es kann durch die farbige Markierung festgestellt werden, dass sich das DL-Modell auf den Bereich des Tumors konzentriert, wobei der Rand des Tumors eine besonders hohe Aktivierung aufweist.
Bei Bild C und D ist ein MDM2-negatives Lipom dargestellt, das das Modell korrekt und mit einer Sicherheit von 64 % klassifiziert hat (von Schacky et al., 2023; Status: eingereicht).

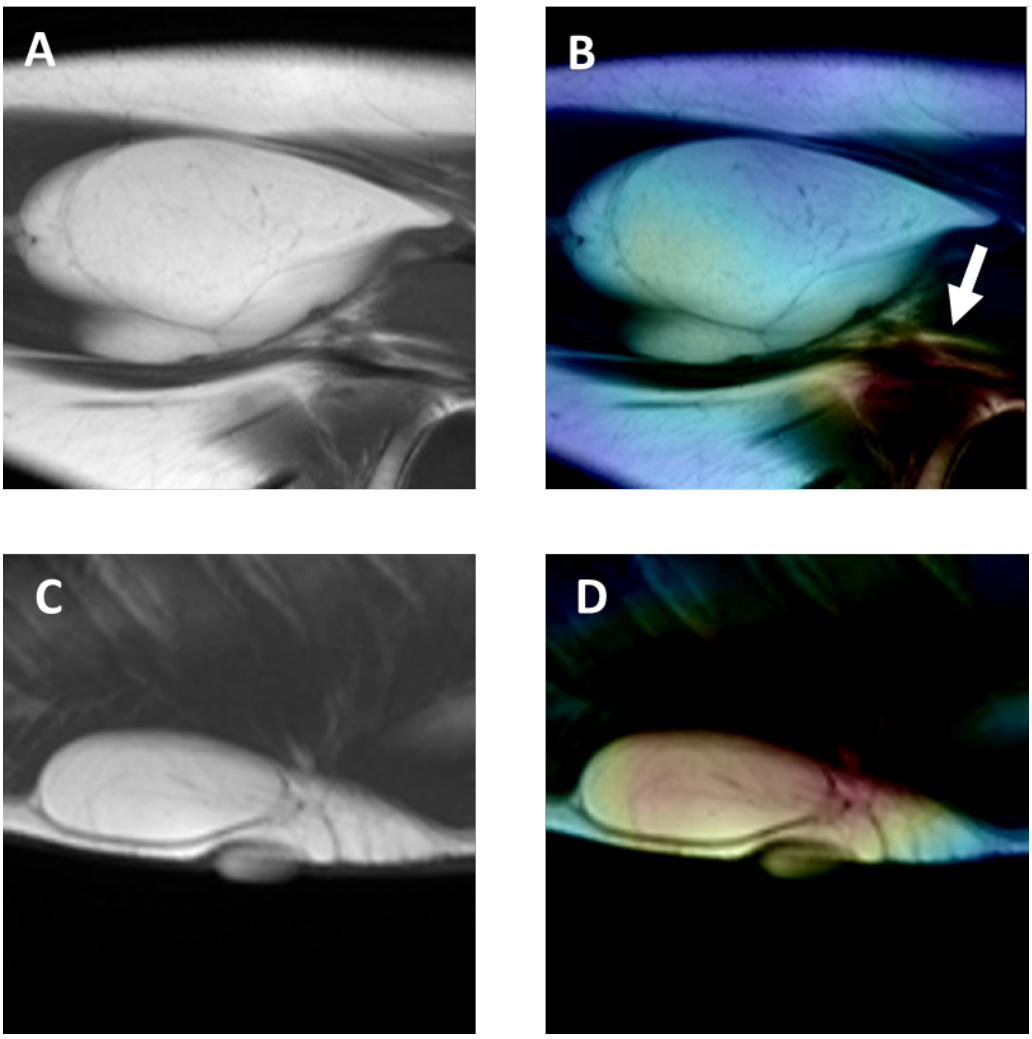


Abbildung 16: Hier sind zwei Beispiele dargestellt, in denen der DL-Algorithmus inkorrekt klassifiziert hat.
Die rote Färbung stellt eine hohe Aktivierung dar und die blaue eine niedrige. Bei beiden Bildern handelt es sich um T1-gewichtete MR-Bilder. Die Bilder B und D stellen jeweils die von einer Grad-CAM überlagerten Input-Bilder des vorhergehenden Bildes dar.
Bei Bild A und Bild B handelt es sich um einen MDM2-positiven ALT der vom 2D-DL-Algorithmus mit einer sehr niedrigen Sicherheit von 55 % fälschlich als Lipom eingestuft wurde. Wie in der Grad-CAM zu sehen ist, konzentriert sich das DL-Modell auf Bereiche außerhalb des Tumors. Der weiße Pfeil in Bild B deutet auf einen Bereich der höchsten Aktivierung außerhalb des Tumors.
Bei Bild B und Bild C wird ein MDM2-negatives Lipom dargestellt, das vom 2D-DL-Modell fälschlich als ALT eingestuft wurde. Auch in diesem Fall war die Sicherheit, mit der der Algorithmus das ALT einstufte, bei sehr niedrigen 53 % (von Schacky et al., 2023; Status: eingereicht).

# 5. Diskussion

In dieser Studie wurde ein DL-Algorithmus zur Differenzierung von ALTs von Lipomen erstellt. Diese Performanz wurde dabei mit der der Radiologen verglichen. Hierzu wurden 2D- und 3D-Modelle entwickelt, die auf MR-Bildern in T1w, T2w, und T1fsgd trainiert und angewendet wurden. Die besten Leistungen wurden mit einem 2D-Modell erzielt, das mit einer Genauigkeit von 95 % bei einer Sensitivität von 89 % und einer Spezifität von 100 % mit einer AUC von 0,98 arbeitete.

Es gelang dem DL-Modell, die Performanz von Assistenzärzt*innen für Radiologie zu übertreffen und lieferte vergleichbare Performanz wie Fachärzt*innen für Radiologie.
Nach den Informationen unserer Forschungsgruppe ist dies eine der ersten Studien, die sich damit beschäftigt, einen DL-Ansatz zur Unterscheidung von ALTs und Lipomen auf MR-Bildern zu entwickeln. Durch die Forschung von (Thornhill et al., 2014) wurden bereits Tumoranalysen anhand von Textur und Form durchgeführt, um 20 Liposarkome von 24 Lipomen zu unterscheiden. Die erreichte Genauigkeit lag in dieser früheren Studie bei 89 %: Sie wies eine Sensitivität von 88 % sowie eine Spezifität von 90 % auf. Damit übertraf der hier angewendete Algorithmus die Ergebnisse von Radiolog*innen mit vier und neun Jahren Erfahrung. Die Radiolog*innen kamen auf eine Genauigkeit von 78 %, eine Sensitivität von 77 % und eine Spezifität von 79 % mit einer pathologischen Analyse als Referenzstandard.
Die Problematik dieser Studie liegt in der kleinen Stichprobengröße sowie der verschiedenen Subtypen der zur Analyse herangezogenen Liposarkome. Weiterhin wurde bei dieser Studie kein MDM2-Amplifikationsstatus erhoben, wodurch die Studie anfällig für eine falsche Klassifikation von Lipomen als ALTs ist (Nagano et al., 2015).

In einer Forschungsarbeit von Vos et al. wurde, wie bei der vorliegenden Forschungsarbeit, der Amplifikationsstatus als Referenzstandard herangezogen (Vos et al., 2019). Der hier verwendete Radiomics-Ansatz, der auf manueller Extraktion von Merkmalen beruht, erreicht nicht die Leistung des in dieser For-

schungsarbeit verwendeten DL-Modells. Es wurden 58 Patienten mit einem Lipom und 58 Patienten mit WDL aus 41 verschiedenen MRT-Bildgebungen verwendet. Hier wurden T1w- und T2w-MR-Bilddateien benutzt und auch die Expertise von Radiologen zum Vergleich der Performanz in die Studie miteingeschlossen. Diese waren hinsichtlich des Alters und des Geschlechts der Patienten allerdings nicht verblindet, da der Radiomics-Ansatz auch Daten außerhalb der Bilddaten nutzt.
Das Radiomics-Modell, das ausschließlich T1w-MR-Bilder verwendete, erreichte eine Genauigkeit von 67 %, eine Sensitivität von 60 % und eine Spezifität von 74 % mit einer AUC von 0,69.
Das Radiomics-Modell, das sowohl die T1w als auch die T2w berücksichtigte, erreichte mit einer Genauigkeit von 75 %, einer Sensitivität von 66 % und einer Spezifität von 84 % ein etwas besseres Resultat als das reine T1w-MR-Modell.
Beim Vergleich wird deutlich, dass der in dieser Forschungsarbeit verwendete Ansatz eines DL-Algorithmus eine bessere Leistung erzielt als der Radiomics-Ansatz in der Studie von Vos et al. von 2019.
Die bessere Performance von DL-Modellen im Vergleich zu Radiomics-Modellen zeigt sich auch in anderen Bereichen, wie beispielsweise bei der Vorhersage von axillären Lymphknotenmetastasen bei Brustkrebs unter Verwendung von Ultraschallbildern (Sun et al., 2020).
Auch bei der Differenzierung gutartiger von bösartigen Läsionen der Brust konnte gezeigt werden, dass die Verwendung eines DL-Ansatzes ein signifikant besseres Ergebnis erzielen konnte. Außerdem ist zu beachten, dass ein größerer Satz an Trainingsdaten die Performanz eines DL-Modells weiter verbessern kann, die des Radiomics-Modells allerdings nicht (Truhn et al., 2019).

Bei dem in Zukunft zu erwartenden starken Anstieg an verfügbaren Daten und der Rechenleistung, wird sich der Abstand des DL-Modells zu anderen Ansätzen weiter vergrößern, da sich bemerkenswerte Fähigkeiten für die Bildanalyse bieten. Das Interesse am Einsatz dieser Technologie in der Forschung und Anwendung ist dementsprechend groß. Die Anzahl der Publikationen auf diesem Gebiet nehmen stark zu und umfassen bereits alle wichtigen Organsysteme.
In den kommenden Jahren wird erwartet, dass sich DL-Modelle vom mehrheitlichen Einsatz in der Forschung zunächst als Hilfsmittel für eine ganz konkrete

Unterscheidung oder Klassifizierung und anschließend im klinischen Alltag der Radiologie durchsetzen werden (Soffer et al., 2019).
Später wird erwartet, dass die DL-Algorithmen einen ganzheitlicheren Ansatz verfolgen werden, bei dem sie gleichzeitig mehrere Differenzierungen und Klassifizierungen gleichzeitig vorgenommen werden können und der Algorithmus eine automatische Analyse vollzieht. Eine komplett automatische Läsionserkennung und Klassifikation in der Mammografie wurden bereits erstellt und evaluiert (Carneiro et al., 2017; Dhungel et al., 2017).
Sehr vielversprechend sind auch Ansätze, bei denen außer Bildinformationen auch weitere Daten in das DL-Modell eingegeben werden. So wurden bei der Untersuchung der Performanz eines DL-Algorithmus beim Staging der Leberfibrose außer Bilddaten noch zusätzliche Daten eingegeben, wie beispielsweise positive oder negative Testergebnisse von Hepatitis-B und Hepatitis-C Tests. Dies verbesserte die Performanz des DL-Algorithmus noch weiter (Yasaka et al., 2018).
In der hier vorliegenden Studie konnte gezeigt werden, dass der MDM2-Amplifikations-Status mithilfe eines DL-Modells mit einer hohen AUC vorhergesagt werden konnte. Dies könnte in der Zukunft dazu führen, dass bei entsprechender Weiterentwicklung die molekularen Subtypen eines Weichteiltumors mithilfe von MR-Bilddateien bestimmt werden könnten. In früheren Studien konnte gezeigt werden, dass die Septierung und die Aufnahme von Kontrastmittel wichtige Anhaltspunkte für die Unterscheidung von Lipomen und ALTs darstellen. Auch bei dem DL-Modell dieser Studie konnte durch Anwendung von Grad-CAMs gezeigt werden, dass diese Merkmale für das DL-Modell eine hohe Relevanz für die Entscheidung haben (Knebel et al., 2019; Nardo et al., 2020).

Zu den Limitationen dieser Arbeit gehört, dass hier mit einem retrospektiven Studiendesign gearbeitet wurde, was immer mögliche Verzerrungen bei der Auswahl des Studienkollektivs verbunden ist. Es wurde außerdem mit einer begrenzten Anzahl an Patienten gearbeitet, die am Institut für diagnostische und interventionelle Radiologie im Klinikum rechts der Isar vom Januar 2010 bis zum Oktober 2018. Eine größere Patientenanzahl kann hier für ein noch besser trainierten DL-Algorithmus sorgen. Auch wurde in dieser Studie nur die Differenzierung zwi-

schen ALTs und Lipomen betrachtet, andere Weichteilerkrankungen wurden ausgeschlossen. Die ausschließliche Nutzung von Patientendaten des Institutes für diagnostische und interventionelle Radiologie am Klinikum rechts der Isar, kann über die Anwendbarkeit des verwendeten DL-Algorithmus in anderen Instituten nur eingeschränkt Aufschluss geben. Die Anwendung anderer MR-Geräte kann die Erkennung reduzieren und einen neuen, entsprechend angepassten Trainingsdatensatz erfordern.

Zusammenfassend lässt sich sagen, dass die 2D- und 3D-Modelle bei dieser Studie in der Lage waren zuverlässig zwischen ALTs und Lipomen zu unterscheiden. Der DL-Algorithmus übertraf die Leistung von Assistenzärzt*innen und Fachärzt*innen und war vergleichbar mit der Leistung von Oberärzt*innen mit Spezialisierung in muskuloskelettaler Radiologie. Insbesondere in Krankenhäusern ohne hochspezialisierte Radiolog*innen in diesem Gebiet wäre die Einführung einer solchen Technologie eine große Hilfe für die Differenzierung von Lipomen und ALTs. Systeme zur automatisierten Beurteilung von Hautveränderungen etwa können bereits heute “den Facharzt zum Patienten bringen” (Esteva et al., 2017). KI-Systeme werden daher wahrscheinlich in der klinischen Routine der tumororthopädischen Radiologie sowohl spezialisierte als auch nicht-spezialisierte Radiolog*innen und (Fach-) Ärzt*innen aus anderen Disziplinen bei bildgebenden diagnostischen und therapeutischen Entscheidungen unterstützen.

# 6. Zusammenfassung

## 6.1 Zusammenfassung auf Deutsch

In der Forschungsarbeit „Entwicklung von 2D- und 3D-Deep-Learning-Ansätzen zur Differenzierung von Atypischen Lipomatösen Tumoren von Lipomen basierend auf magnetresonanztomografischen Bildern“ wurde eine retrospektive Studie durchgeführt, die die Anwendung von Deep-Learning-Algorithmen zur Unterscheidung zwischen atypischen lipomatösen Tumoren (ALTs) und Lipomen in der Bildgebung von Patienten mit lipomatösen Neoplasien der oberen Extremität, der unteren Extremität oder des Rumpfes untersucht. Die Patientenauswahl erfolgte anhand von Daten aus dem Klinikum rechts der Isar zwischen Januar 2010 und Oktober 2018. Insgesamt wurden 260 Patienten in die Studie aufgenommen, von denen 98 histopathologisch auf ihren MDM2-Status hin untersucht wurden. Dabei wurde kein Fall von WDL festgestellt, und nur Patienten mit ALTs wurden in die Studie eingeschlossen.

Die Bildgebung erfolgte vor der Operation mithilfe von MRT-Scannern mit 1,5-Tesla- oder 3-Tesla-Magneten und umfasste verschiedene Sequenzen, darunter T1-gewichtete, T2-gewichtete und T1-gewichtete Sequenzen nach Gabe von Kontrastmitteln (T1w fsgd). Die Segmentierung der Tumore wurde manuell von Experten unter Verwendung von Open-Source-Software durchgeführt.

Zur Unterscheidung zwischen ALTs und Lipomen wurden sowohl 2D- als auch 3D-Deep-Learning-Modelle entwickelt und trainiert. Die 2D-Modelle basierten auf einem in ImageNET vortrainierten Dense Convolutional Network (DenseNET 201), während die 3D-Modelle auf einer 2D-ResNet-Architektur aufbauten.

Die Ergebnisse der Modelle wurden auf verschiedenen Sequenzen ausgewertet, wobei das 2D-Modell auf T1w-Bildern die höchste AUC (0,98) erzielte. Die 3D-Modelle schnitten etwas schlechter ab, wobei das Modell auf T1w-Bildern eine AUC von 0,91 erreichte.

Die Modelle wurden mit den Ergebnissen von Radiologen verglichen. Dabei zeigte sich, dass das beste DL-Modell die Leistung von Assistenzärzten übertraf und etwa die gleiche Leistung wie Fachärzte für Radiologie erbrachte.

Die Visualisierung der Modelle zeigte, dass die DL-Algorithmen sich hauptsächlich auf den Bereich des Tumors, insbesondere auf den Rand des Tumors, konzentrierten. In einigen Fällen führte die Konzentration auf Bereiche außerhalb des Tumors zu inkorrekten Klassifikationen.

Insgesamt liefert die Studie vielversprechende Ergebnisse zur Anwendung von Deep-Learning-Modellen zur Unterscheidung zwischen ALTs und Lipomen in der MRT-Bildgebung von lipomatösen Neoplasien. Die Modelle zeigten eine hohe Genauigkeit und Sensitivität, was auf ihr Potenzial hinweist, in der klinischen Praxis eingesetzt zu werden.

## 6.2 Zusammenfassung auf Englisch

In the research paper titled "Development of 2D and 3D Deep Learning Approaches for the Differentiation of Atypical Lipomatous Tumors from Lipomas Based on Magnetic Resonance Imaging," a retrospective study was conducted to explore the application of deep learning algorithms in distinguishing between atypical lipomatous tumors (ALTs) and lipomas in the imaging of patients with lipomatous neoplasms of the upper extremity, lower extremity, or trunk. Patient selection was based on data from Klinikum rechts der Isar between January 2010 and October 2018. A total of 98 patients were enrolled in the study, of whom 114 were histopathologically examined for their MDM2 status. No cases of WDL were identified, and only patients with ALTs were included in the study.

Imaging was performed preoperatively using MRI scanners with 1.5-tesla or 3-tesla magnets, encompassing various sequences, including T1-weighted, T2-weighted, and contrast-enhanced T1-weighted sequences. Tumor segmentation was carried out manually by experts using open-source software.

To differentiate between ALTs and lipomas, both 2D and 3D deep learning models were developed and trained. The 2D models were based on a Dense Convolutional Network (DenseNET 201) pretrained on ImageNet, while the 3D models were built on a 2D ResNet architecture.

The results of the models were evaluated across various sequences, with the 2D model achieving the highest AUC (0.98) on T1-weighted images. The 3D models performed slightly less well, with the T1-weighted model achieving an AUC of 0.91.

The models were compared with the results of radiologists, and it was observed that the best deep learning model outperformed the performance of assistant physicians and achieved a performance comparable to that of radiology specialists.

Visualization of the models indicated that the deep learning algorithms primarily focused on the tumor area, particularly the tumor's edges. In some cases, concentrating on areas outside the tumor led to incorrect classifications.

Overall, the study provides promising results regarding the application of deep learning models for distinguishing between ALTs and lipomas in MRI imaging of lipomatous neoplasms. The models exhibited high accuracy and sensitivity, suggesting their potential for use in clinical practice.

# Danksagung

Zunächst möchte ich meinem Doktorvater PD Dr. med. Benedikt J. Schwaiger danken, mich in die Forschungsgruppe aufgenommen zu haben.

PD Dr. med. Alexandra S. Gersing möchte ich für eine sehr produktive Zusammenarbeit danken, für die Vorbereitung auf die Präsentation der Arbeit auf dem ESSR-Kongress sowie für das hilfreiche wiederholte Feedback.

Dr. med. Sarah C. Foreman möchte ich herzlich für die gute Einführung in die Thematik und die außergewöhnlich gute Betreuung und Zusammenarbeit bedanken, sowie für die Aufnahme in die Arbeitsgruppe mit einer so interessanten Promotions-Thematik.

Auch möchte ich mich für die Zusammenarbeit mit Dr. med. Claudio E. von Schacky bedanken, der immer wieder sehr hilfreichen Input gab und mich bei der Präsentation der Forschungsarbeit auf dem Kongress sehr unterstützt hat.

PD Dr. med. Jan Peeken möchte ich dafür danken, mein Betreuer für diese Arbeit gewesen zu sein und stets ein offenes Ohr gehabt zu haben.

Fernando Navarro möchte ich für die produktive Zusammenarbeit danken.

Bei Prof. Dr. med. Carsten Rist möchte ich mich für die vielen hilfreichen Gespräche danken. Auch dafür, dass er mir den Kontakt zu meiner Arbeitsgruppe herstellte und mir somit eine Promotion in einem von mir favorisierten Bereich ermöglicht hat.

Ganz besonders danke ich meinen Eltern, Elvira M. Kramp und Dipl. Ing. Wolfgang G. Kramp, die mir das Studium erst ermöglichten und mich jederzeit unterstützten.

Lisa Schmidhuber möchte ich für die Durchsicht dieser Arbeit danken. Und für die vielen hilfreichen Tipps und Gespräche bei koffeinhaltigen Heißgetränken.

Der gesamten MSK-Forschungsgruppe der TU München möchte ich meinen herzlichsten Dank aussprechen. Ich bedanke mich bei allen, die mich durch einen guten Austausch und Hilfe bei der Entstehung dieser Arbeit unterstützt haben.

# Abbildungs- und Tabellenverzeichnis

# Publikationsliste

Carneiro, G., Nascimento, J., & Bradley, A. P. (2017). Automated Analysis of Unregistered Multi-View Mammograms With Deep Learning. *IEEE Trans Med Imaging*, *36*(11), 2355-2365. https://doi.org/10.1109/tmi.2017.2751523

Claudio E. von Schacky*, F. N., Daniel W. Kramp, Tim Tomov, Jan Neumann, Alexander A. Fingerle, Carolin Knebel, Rüdiger von Eisenhart-Rothe, Florian T. Gassert, Felix G. Gassert, Amelia Jiménez-Sànchez, Katja Specht, Jan S. Kirschke, Benedikt J. Schwaiger, Marcus R. Makowski, Jan C. Peeken, Stephanie E. Combs, Bjoern H. Menze, Klaus Woertler, Sarah C. Foreman†, Alexandra S. Gersing†. (2021). Development of 2D and 3D Deep Learning Approaches to Differentiate Atypical Lipomatous Tumors from Lipomas with Magnetic Resonance Imaging and Comparison with Radiologists. Status des Papers: Eingereicht bei European Radiology, zum jetzigen Zeitpunkt noch nicht veröffentlicht (under revisions) bei European Radiology

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition,

Dhungel, N., Carneiro, G., & Bradley, A. P. (2017). A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Med Image Anal*, *37*, 114-128. https://doi.org/10.1016/j.media.2017.01.009

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

Gassert, F. G., Gassert, F. T., Specht, K., Knebel, C., Lenze, U., Makowski, M. R., von Eisenhart-Rothe, R., Gersing, A. S., & Woertler, K. (2021). Soft tissue masses: distribution of entities and rate of malignancy in small lesions. *BMC Cancer*, *21*(1), 93. https://doi.org/10.1186/s12885-020-07769-2

Goldblum, J. R., Folpe, A. L., & Weiss, S. W. (2014). *Enzinger and Weiss's Soft Tissue Tumors*. Elsevier Saunders. https://books.google.de/books?id=D_5tmwEACAAJ

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. (2018). Artificial intelligence in radiology. *Nat Rev Cancer*, *18*(8), 500-510. https://doi.org/10.1038/s41568-018-0016-5

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Husseini, M., Sekuboyina, A., Loeffler, M., Navarro, F., Menze, B. H., & Kirschke, J. S. (2020). Grading loss: a fracture grade-based metric loss for vertebral fracture detection. International Conference on Medical Image Computing and Computer-Assisted Intervention,

Jiménez-Sánchez, A., Mateus, D., Kirchhoff, S., Kirchhoff, C., Biberthaler, P., Navab, N., Ballester, M. A. G., & Piella, G. (2019). Medical-based deep curriculum learning for improved fracture classification. International Conference on Medical Image Computing and Computer-Assisted Intervention,

Jo, V. Y., & Fletcher, C. D. (2014). WHO classification of soft tissue tumours: an update based on the 2013 (4th) edition. *Pathology*, *46*(2), 95-104. https://doi.org/10.1097/pat.0000000000000050

Kindblom, L.-G., Angervall, L., Stener, B., & Wickbom, I. (1974). Intermuscular and intramuscular lipomas and hibernomas. A clinical, roentgenologic, histologic, and prognostic study of 46 cases [https://doi.org/10.1002/1097-0142(197403)33:3<754::AID-CNCR2820330322>3.0.CO;2-F]. *Cancer*, *33*(3), 754-762. https://doi.org/https://doi.org/10.1002/1097-0142(197403)33:3<754::AID-CNCR2820330322>3.0.CO;2-F

Knebel, C., Neumann, J., Schwaiger, B. J., Karampinos, D. C., Pfeiffer, D., Specht, K., Lenze, U., von Eisenhart-Rothe, R., Rummeny, E. J., Woertler, K., & Gersing, A. S. (2019). Differentiating atypical lipomatous tumors from lipomas with magnetic resonance imaging: a comparison with MDM2 gene amplification status. *BMC Cancer*, *19*(1), 309. https://doi.org/10.1186/s12885-019-5524-5

Kransdorf, M. J., Bancroft, L. W., Peterson, J. J., Murphey, M. D., Foster, W. C., & Temple, H. T. (2002). Imaging of fatty tumors: distinction of lipoma and well-differentiated liposarcoma. *Radiology*, *224*(1), 99-104. https://doi.org/10.1148/radiol.2241011113

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., . . . Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89-94. https://doi.org/10.1038/s41586-019-1799-6

Moll, U. M., & Petrenko, O. (2003). The MDM2-p53 interaction. *Mol Cancer Res*, *1*(14), 1001-1008.

Murphey, M. D., Carroll, J. F., Flemming, D. J., Pope, T. L., Gannon, F. H., & Kransdorf, M. J. (2004). From the archives of the AFIP: benign musculoskeletal lipomatous lesions. *Radiographics*, *24*(5), 1433-1466. https://doi.org/10.1148/rg.245045120

Myhre-Jensen, O. (1981). A consecutive 7-year series of 1331 benign soft tissue tumours. Clinicopathologic data. Comparison with sarcomas. *Acta Orthop Scand*, *52*(3), 287-293. https://doi.org/10.3109/17453678109050105

Nagano, S., Yokouchi, M., Setoguchi, T., Ishidou, Y., Sasaki, H., Shimada, H., & Komiya, S. (2015). Differentiation of lipoma and atypical lipomatous tumor by a scoring system: implication of increased vascularity on pathogenesis of liposarcoma. *BMC Musculoskelet Disord*, *16*, 36. https://doi.org/10.1186/s12891-015-0491-8

Nardo, L., Abdelhafez, Y. G., Acquafredda, F., Schirò, S., Wong, A. L., Sarohia, D., Maroldi, R., Darrow, M. A., Guindani, M., Lee, S., Zhang, M., Moawad, A. W., Elsayes, K. M., Badawi, R. D., & Link, T. M. (2020). Qualitative evaluation of MRI features of lipoma and atypical lipomatous tumor: results from a multicenter study. *Skeletal Radiol*, *49*(6), 1005-1014. https://doi.org/10.1007/s00256-020-03372-5

Navarro, F., Conjeti, S., Tombari, F., & Navab, N. (2018). Webly supervised learning for skin lesion classification. International Conference on Medical Image Computing and Computer-Assisted Intervention,

Navarro, F., Sekuboyina, A., Waldmannstetter, D., Peeken, J. C., Combs, S. E., & Menze, B. H. (2020). Deep reinforcement learning for organ localization in CT. Medical Imaging with Deep Learning,

Navarro, F., Shit, S., Ezhov, I., Paetzold, J., Gafita, A., Peeken, J. C., Combs, S. E., & Menze, B. H. (2019). Shape-aware complementary-task learning for multi-organ

segmentation. International Workshop on Machine Learning in Medical Imaging,

Norman, B., Pedoia, V., & Majumdar, S. (2018). Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. *Radiology*, *288*(1), 177-185. https://doi.org/10.1148/radiol.2018172322

O'Donnell, P. W., Griffin, A. M., Eward, W. C., Sternheim, A., White, L. M., Wunder, J. S., & Ferguson, P. C. (2013). Can Experienced Observers Differentiate between Lipoma and Well-Differentiated Liposarcoma Using Only MRI? *Sarcoma*, *2013*, 982784. https://doi.org/10.1155/2013/982784

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 1-8.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention,

Ryan, S., Visgauss, J., Kerr, D., Helmkamp, J., Said, N., Vinson, E., O'Donnell, P., Li, X., Jung, S. H., Cardona, D., Eward, W., & Brigman, B. (2018). The Value of MRI in Distinguishing Subtypes of Lipomatous Extremity Tumors Needs Reassessment in the Era of MDM2 and CDK4 Testing. *Sarcoma*, *2018*, 1901896. https://doi.org/10.1155/2018/1901896

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision,

Soffer, S., Ben-Cohen, A., Shimon, O., Amitai, M. M., Greenspan, H., & Klang, E. (2019). Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide. *Radiology*, *290*(3), 590-606. https://doi.org/10.1148/radiol.2018180547

Subramanian, V. (2018). *Deep Learning with PyTorch: A practical approach to building neural network models using PyTorch*. Packt Publishing Ltd.

Sun, Q., Lin, X., Zhao, Y., Li, L., Yan, K., Liang, D., Sun, D., & Li, Z.-C. (2020). Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region [Original Research]. *Frontiers in Oncology*, *10*(53). https://doi.org/10.3389/fonc.2020.00053

Tedesco, N. S., & Henshaw, R. M. (2016). Unplanned Resection of Sarcoma. *J Am Acad Orthop Surg*, *24*(3), 150-159. https://doi.org/10.5435/jaaos-d-15-00074

Thomson, A. (2010). Neocortical layer 6, a review [Review]. *Frontiers in Neuroanatomy*, *4*(13). https://doi.org/10.3389/fnana.2010.00013

Thornhill, R. E., Golfam, M., Sheikh, A., Cron, G. O., White, E. A., Werier, J., Schweitzer, M. E., & Di Primio, G. (2014). Differentiation of lipoma from liposarcoma on MRI using texture and shape analysis. *Academic radiology*, *21*(9), 1185-1194.

Truhn, D., Schrading, S., Haarburger, C., Schneider, H., Merhof, D., & Kuhl, C. (2019). Radiomic versus Convolutional Neural Networks Analysis for Classification of

Contrast-enhancing Lesions at Multiparametric Breast MRI. *Radiology*, *290*(2), 290-297. https://doi.org/10.1148/radiol.2018181352

von Schacky, C. E., Navarro, F., Kramp, D. W., Tomov, T., Neumann, J., Fingerle, A. A., Knebel, C., von Eisenhart-Rothe, R., Gassert, F. T., Gassert, F. G., Jiménez-Sánchez, A., Specht, K., Kirschke, J. S., Schwaiger, B. J., Makowski, M. R., Peeken, J. C., Combs, S. E., Menze, B. H., Woertler, K., . . . Gersing, A. S. (2023). Development of a Deep Learning Approach to Differentiate Atypical Lipomatous

Tumors from Lipomas with Magnetic Resonance Imaging and Comparison with Radiologists [jurnal article]. *submitted to European Radiology*.

von Schacky, C. E., Sohn, J. H., Liu, F., Ozhinsky, E., Jungmann, P. M., Nardo, L., Posadzy, M., Foreman, S. C., Nevitt, M. C., Link, T. M., & Pedoia, V. (2020). Development and Validation of a Multitask Deep Learning Model for Severity Grading of Hip Osteoarthritis Features on Radiographs. *Radiology*, *295*(1), 136-145. https://doi.org/10.1148/radiol.2020190925

Vos, M., Starmans, M. P. A., Timbergen, M. J. M., van der Voort, S. R., Padmos, G. A., Kessels, W., Niessen, W. J., van Leenders, G., Grünhagen, D. J., Sleijfer, S., Verhoef, C., Klein, S., & Visser, J. J. (2019). Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI. *The British journal of surgery*, *106*(13), 1800-1809. https://doi.org/10.1002/bjs.11410

Wu, J. S., & Hochman, M. G. (2009). Soft-tissue tumors and tumorlike lesions: a systematic imaging approach. *Radiology*, *253*(2), 297-316. https://doi.org/10.1148/radiol.2532081199

Xu, X., Zhou, F., Liu, B., Fu, D., & Bai, X. (2019). Efficient multiple organ localization in CT image using 3D region proposal network. *IEEE transactions on medical imaging*, *38*(8), 1885-1898.

Yasaka, K., Akai, H., Kunimatsu, A., Abe, O., & Kiryu, S. (2018). Liver Fibrosis: Deep Convolutional Neural Network for Staging by Using Gadoxetic Acid-enhanced Hepatobiliary Phase MR Images. *Radiology*, *287*(1), 146-155. https://doi.org/10.1148/radiol.2017171928

Anmerkung zur Quelle "*Claudio E. von Schacky*, F. N., Daniel W. Kramp, Tim Tomov, Jan Neumann, Alexander A. Fingerle, Carolin Knebel, Rüdiger von Eisenhart-Rothe, Florian T. Gassert, Felix G. Gassert, Amelia Jiménez-Sànchez, Katja Specht, Jan S. Kirschke, Benedikt J. Schwaiger, Marcus R. Makowski, Jan C. Peeken, Stephanie E. Combs, Bjoern H. Menze, Klaus Woertler, Sarah C. Foreman†, Alexandra S. Gersing†. (2021). Development of 2D and 3D Deep Learning Approaches to Differentiate Atypical Lipomatous Tumors from Lipomas with Magnetic Resonance Imaging and Comparison with Radiologists.*"

Das Paper wurde eingereicht und befindet sich derzeit unter dem genannten Titel im Status „under revisions" bei European Radiology.