Technische Universität München
TUM School of Medicine and Health

Differenzierung von osteoporotischen und pathologischen Wirbelkörperfrakturen durch Deep Learning Analysen

Anna-Sophia Walburga Dietrich

Vollständiger Abdruck der von der TUM School of Medicine and Health der Technischen Universität München zur Erlangung einer Doktorin der Medizin (Dr. med.) genehmigten Dissertation.

Vorsitz: apl. Prof. Dr. Bernhard Haslinger

Prüfende der Dissertation:
1. Priv.-Doz. Dr. Alexandra Gersing
2. Priv.-Doz. Dr. Benedikt Schwaiger

Die Dissertation wurde am 12.02.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Medicine and Health am 07.08.2024 angenommen.

Inhaltsverzeichnis

1. Einleitung

1.1 Künstliche Intelligenz und Deep Learning Algorithmen

1.1.1 Überblick

Künstliche Intelligenz (KI) hat in den vergangenen Jahren in wissenschaftlichen Kreisen viel Aufmerksamkeit erregt. Dies wird durch die jährlich steigende Zahl, der über dieses Thema veröffentlichten, wissenschaftlichen Arbeiten, gezeigt. Die Möglichkeiten und Chancen, die KI besonders im medizinischen Bereich bietet, werden derzeit von vielen Wissenschaftlern und wissenschaftlich tätigen Ärzten untersucht (Langs et al., 2020; McBee et al., 2018; Sahiner et al., 2019; Zaharchuk et al., 2018).

Eine Unterform der KI ist das maschinelle Lernen. Darunter versteht man die Entwicklung von Algorithmen, die es Computern ermöglicht, selbstständig aus vorhandenen Daten zu lernen, ohne dass sie dafür explizit programmiert werden mussten. Maschinelles Lernen kann wiederum in zwei Untergruppen unterteilt werden: Überwachtes und unüberwachtes Lernen (Langs et al., 2020; McBee et al., 2018; Zaharchuk et al., 2018).

Unter überwachtem Lernen versteht man das Trainieren von Algorithmen mithilfe von annotierten Datensätzen. Diese Datensätze bestehen aus Paaren von Eingaben und entsprechenden Ausgaben. Deep Learning Algorithmen gehören zur Gruppe des überwachten Lernens. (Langs et al., 2020; Zaharchuk et al., 2018).

Im Gegensatz dazu steht das unüberwachte Lernen. Beim unüberwachten Lernen wird der Algorithmus mithilfe von nicht beschrifteten oder annotierten Datensätzen, also ohne menschliche Hilfe, trainiert (Langs et al., 2020; Zaharchuk et al., 2018).

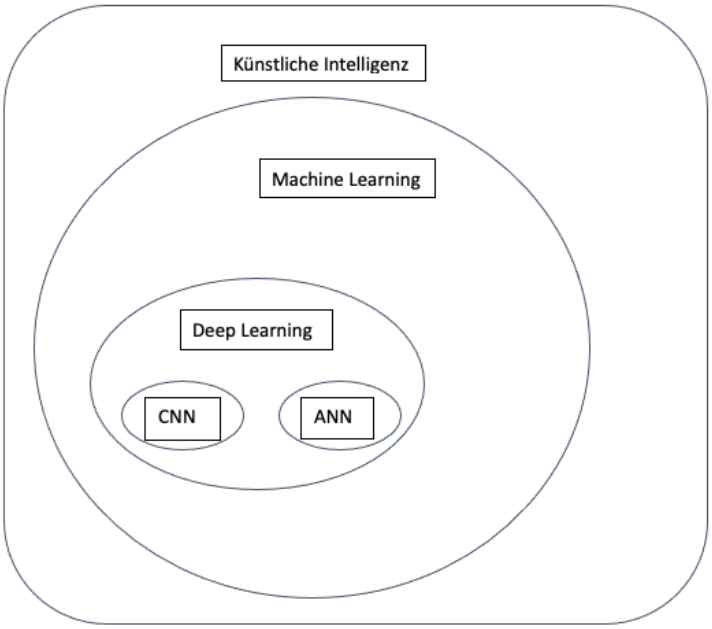

Abbildung 1: Schematische Übersicht über Künstliche Intelligenz.

1.1.2 Deep Learning Algorithmen und Netzwerkarchitekturen

Deep Learning Algorithmen basieren auf neuronalen Netzwerken, genauer gesagt aus Schichten von künstlichen Neuronen. Jede Schicht enthält eine Anzahl von Einheiten, wobei jede Einheit eine vereinfachte Darstellung eines Neurons ist. Die Architektur dieser neuronalen Netzwerke kann vereinfacht mit der Architektur der Neuronen im menschlichen Gehirn verglichen werden (Matsumura et al., 2019). Im Zusammenhang mit Deep Learning bedeutet "Deep" nicht ein "tieferes Verständnis", sondern bezieht sich auf die "Tiefe" des Netzwerks. In der Regel bestehen diese Netzwerke aus zehn bis 30 Schichten, sie können aber auch aus mehreren 100 Schichten bestehen (Mazurowski et al., 2019).

Eine Form der neuronalen Netzwerke sind Artificial neural networks (ANNs), auch künstliche neuronale Netze genannt (KNNs) genannt. Diese haben sich bereits als sehr vielversprechend gezeigt. ANNs sind vollständig verbundene Datenverarbeitungsnetze,

die aus mehreren Ebenen von Neuronen bestehen. Zwischen der Eingabe- und der Ausgabeschicht befinden sich in der Regel mehrere verborgene Schichten, die analog zu den Interneuronen im menschlichen Gehirn agieren. In diesen verborgenen Zwischenschichten werden nichtlineare Transformationen durchgeführt, hier findet das eigentliche „Lernen“ statt. In der Regel befinden sich in den verborgenen Schichten mehr Neuronen als in den Eingabe- oder Ausgabeschichten. Die Fragestellung definiert die Ausgabeschichten. Soll eine kategorische Fragestellung mit zwei Kategorien, zum Beispiel „Ja“ oder „Nein“ beantwortet werden, werden zwei endgültige Ausgabeschichten benötigt. Gäbe es drei Kategorien, würden drei endgültige Ausgabeschichten benötigt werden (Zaharchuk et al., 2018).

Die Netzwerkarchitektur beschreibt wie viele Schichten es gibt und wie viele Neuronen pro Schicht verwendet werden. Jedes Neuron hat einen numerischen Wert und jede Verbindung zwischen Neuronen hat ein bestimmtes Gewicht. Die Gewichte verbinden Neuronen in verschiedenen Ebenen und geben an, wie stark diese Verbindungen sind (Zaharchuk et al., 2018). Die Gewichte werden mithilfe von Lernstrategien so angepasst, dass je nach Aufgabe mehr oder weniger Wert auf bestimmte Datenpunkte gelegt wird. Geht man von einem Bild mit einer typischen Matrixgröße aus, ist die Anzahl der Gewichte in einem vollständig verknüpften neuronalen Netz immens (256 x 256 = 65.536 Voxel). In diesem Beispiel werden für eine vollständig verknüpfte Schicht mehr als 4 Milliarden Gewichte benötigt (Zaharchuk et al., 2018).

Da diese Netze aufgrund der schieren Anzahl von Gewichten rechenintensiv sind, wird moderne Computerhardware wie Grafikprozessoren (GPUs) benötigt, die ursprünglich in Videospielen eingesetzt wurden. Die Verwendung dieser GPUs ist eine neuere

Entwicklung bei Machine Learning (ML) Algorithmen und hat zu einer höheren Geschwindigkeit und Leistungsfähigkeit von Deep Learning (DL) Algorithmen geführt. Dadurch konnten große Fortschritte bei Bildsegmentierungen, sowie Objekt- und Mustererkennungen gemacht werden (McBee et al., 2018).

Eine weitere Form von neuronalen Netzwerken sind Convolutional Neural Networks (CNNs), die ebenfalls grob der Architektur des menschlichen Gehirns nachempfunden sind (Zaharchuk et al., 2018). CNNs sind im Gegensatz zu ANNs keine vollständig verbundenen Netzwerke (Mazurowski et al., 2019; Zaharchuk et al., 2018).

CNNs stellen die gängigste Architektur für maschinelles Lernen auf Bildern dar. Sie sind eine neuere Entwicklung auf dem Gebiet des maschinellen Lernens und ermöglichen eine robuste, selbstlernende Verarbeitung und Analyse medizinischer Bilder. Die Stärke von CNNs liegt in ihrer Fähigkeit wichtige Merkmale in Bildern zu identifizieren und Datensätze automatisch zu klassifizieren (Zaharchuk et al., 2018). Sie sind deshalb eine sehr leistungsfähige und überzeugende Option der Bildanalyse (Langs et al., 2020). Die Verwendung von CNNs hat sich in einem großen Teil der Forschung auf dem Gebiet der medizinischen Bildgebung durchgesetzt (Mazurowski et al., 2019; Zaharchuk et al., 2018).

CNNs verwenden Kernels. Kernels sind Filter, die verwendet werden, um Merkmale aus den Bilddaten zu extrahieren. In einem ersten Schritt werden in dem untersuchten Bild kleinere Strukturen wie Eckpunkte und Kanten erkannt. Diese werden dann anschließend zu komplexeren Strukturen zusammengefügt (Langs et al., 2020). Diese Kernels werden an jeder Bildposition angewendet, um den Wert des Neurons in der nächsten Schicht zu bestimmen. Zwischen zwei Schichten werden nur die Gewichte für

den Kernel benötigt, der dann über das Bild gerastert wird, um die nächste Schicht zu erhalten. Dadurch kann die Anzahl der Gewichte erheblich reduziert werden. Außerdem wird so räumliche Invarianz ermöglicht: Bildmerkmale können unabhängig von ihrer genauen Position im Bild identifiziert werden (Zaharchuk et al., 2018). Bevor der effiziente Einsatz von CNNs möglich war, mussten diese Merkmale von Hand oder durch weniger leistungsfähige maschinelle Lernmodelle erstellt werden (Lundervold & Lundervold, 2019).

Kompositionalität ist von grundlegender Bedeutung bei der Arbeit mit komplexen Mustern, wie sie in der Radiologie vorkommen. Kompositionalität beschreibt die Tatsache, dass viele Objekte eine Zusammensetzung aus anderen Objekten sind. Langs et al. geben das folgende Beispiel: Wenn sich auf zwei Tischen verschiedene Objekte befinden, gehören die Tische dennoch nicht zu zwei grundsätzlich verschiedenen Kategorien. Diese Unterteilbarkeit wird von einer Lernstrategie genutzt: Die Tische und die jeweiligen Objekte werden als solche erkannt und können voneinander unterschieden werden. Wendet man dies auf bildgebende Verfahren in der Medizin an, kann das, was man beispielsweise auf einem Computertomographie (CT)-Scan sieht, in die einzelnen anatomischen Strukturen zerlegt und ihre räumliche Beziehung zueinander in den Bilddaten bestimmt werden (Langs et al., 2020). Bislang sieht die Leistung von CNNs bei einer Vielzahl von Aufgaben, darunter Objekterkennung, Segmentierung und Klassifizierung von Bilddaten, hervorragend aus. Heute sind Deep Learning Algorithmen in der Lage konkrete Fragestellungen genauso erfolgreich wie ein Mensch zu beantworten oder menschliche Fähigkeiten sogar teilweise zu übertreffen (Hirschmann

et al., 2019; Kalmet et al., 2020; Mnih et al., 2015; Moravčík et al., 2017; Zaharchuk et al., 2018).

1.1.3 Trainieren von Modellen

Deep Learning Algorithmen erstellen Prognosemodelle, z. B. die Wahrscheinlichkeit, dass eine Fraktur mit einem bestimmten Aussehen auf eine osteoporotische Veränderung der Knochenstruktur zurückzuführen ist. Vor allem der Umfang und die Repräsentativität der Daten, mit denen der Algorithmus trainiert wird, ist für die Genauigkeit und den späteren Erfolg des Modells entscheidend (Langs et al., 2020). In den meisten Fällen werden die Algorithmen mit CT- und Magnetresonanztomographie (MRT)-Datensätzen und mit zugewiesenen Zielwerten trainiert, z. B. "pathologische Fraktur" oder "osteoporotische Fraktur". Erkannte Beobachtungen, z. B. die Form oder Größe einer Läsion, werden in Merkmalsvektoren übersetzt. Es werden nicht nur einzelne Merkmale verwendet, sondern gesamte Merkmalsvektoren. So fließen die Beziehungen zwischen den Merkmalen in die Vorhersage ein (Langs et al., 2018). Je größer die Anzahl der Fälle ist, mit denen die neuronalen Netze trainiert werden, desto besser ist das zu erwartende Ergebnis. Dieses theoretische Konzept kann auch in der Realität beobachtet werden: Beispielsweise war ein mit 2D-Daten trainierter Algorithmus zur vollautomatischen Segmentierung des Herzens zur Beurteilung der Herzfunktion erfolgreicher als ein mit 3D-Daten trainierter Algorithmus. Als Erklärungsansatz wurde die schlechtere Verfügbarkeit von 3D-Netzen angeführt. Es waren also weniger Daten vorhanden, mit denen der Datensatz trainiert werden konnte (Baumgartner et al., 2018). Weiterhin wurde auf den begrenzten GPU Speicher hingewiesen, deshalb war

möglicherweise ein Downsampling der Daten erforderlich, was einen Datenverlust mit sich brachte (Baumgartner et al., 2018).

Im ersten Schritt des Trainierens von Modellen wird die Gesamtzahl der Fälle in mehrere Gruppen unterteilt. In der Regel wird der größte Teil, etwa 50-60 %, zum Trainieren des Modells verwendet, weitere 10-20 % zur Validierung und 20-40 % zum Testen. Grundsätzlich kann zwischen kategorischen, z.B. zur Bestimmung der Art der Fraktur, und skalaren, z.B. zur Bestimmung des Krankheitsstadiums, Variablen unterschieden werden (Langs et al., 2020).

Jeder Algorithmus benötigt eine Zielfunktion. Der Algorithmus versucht, diese Zielfunktion zu optimieren. Meist wird ein Minimierungsproblem formuliert, das den Algorithmus veranlasst, die Modellparameter so zu verändern, dass der Fehler, der zwischen der Vorhersage und der erwarteten Aussage liegt, mit jedem Durchlauf kleiner wird. Es werden also mithilfe mathematischer Berechnungen die Ein- und Ausgaben der einzelnen Neurone gewichtet und die Verarbeitung der Informationen optimiert, um im Endeffekt das gewünschte Ergebnis, also den geringsten Fehler zu erlangen (Olczak et al., 2017). Die Daten durchlaufen den Algorithmus mehrmals in einer Schleife, auch Epochen genannt, und mit jedem Durchlauf wird die Genauigkeit des Modells erhöht, indem die Vorhersagen mit der Realität abgeglichen werden, wodurch ein Lernprozess ausgelöst wird (Zaharchuk et al., 2018). Durch dieses mehrmalige Wiederholen des Modells mit unterschiedlichen Daten und mehreren Durchläufen wird das Modell trainiert und die Gewichte verbessert. Nach der Optimierung anderer Parameter, wie der Lernrate und der Anzahl der Epochen, wird der Testsatz verwendet, um die Genauigkeit des Modells zu bewerten. Da die Daten aus diesem Satz dem Modell unbekannt sind, kann damit

abgeschätzt werden, wie gut und erfolgreich das Modell letztendlich auf realen Daten angewendet werden kann. Das zeitintensive Training zahlt sich langfristig aus, da das trainierte Modell bei neuen Daten Zeit spart (Zaharchuk et al., 2018).

1.1.4 Technische Herausforderungen und Lösungsansätze

Overfitting

Wie bereits beschrieben, haben Deep Learning Modelle in der Regel Millionen Gewichte. Es wäre theoretisch möglich, den Zusammenhang zwischen Eingabe- und Ausgabezuständen perfekt darzustellen, wenn man die Anzahl der Variablen in den Berechnungen stark erhöht. Das Problem dabei ist jedoch, dass der Algorithmus zwar perfekt an die Trainingsdaten angepasst ist, aber nicht auf neue Daten verallgemeinert werden kann. Dieses Problem wird als Overfitting bezeichnet. Im Allgemeinen ist der Hauptgrund dafür eine zu geringe Größe des Trainingsdatensatzes. Aus diesem Grund hilft grundsätzlich jede Lösung, mit der die Datenmenge vergrößert werden kann, mit der der Algorithmus trainiert wird, gegen das Overfitting-Problem (Hesamian et al., 2019). Die einfachste Option ist ein größerer Datensatz. Eine andere Möglichkeit ist die Datenaugmentation, auch Datenerweiterung genannt. Darunter versteht man verschiedene Techniken, die den vorhandenen Datensatz vergrößern. Beispielsweise Verschieben, Verdrehen, Verzerren oder Spiegeln von Bilddaten. Da im Regelfall Bilddaten auch dann erkennbar sein sollten, wenn sie in dieser Art und Weise manipuliert wurden, werden solche Techniken häufig und regelmäßig durchgeführt. Obwohl diese Bildmodifikationen nicht mehr Daten hinzufügen, verbessern sie nachweislich die

Robustheit der Modelle, da sie möglicherweise verhindern, dass das Modell Merkmale lernt, die nur in einer bestimmten Ausrichtung auftreten (Srivastava et al., 2014).

Ein weiterer Lösungsansatz ist die Dropout-Technik, die zu einer verbesserten Generalisierung führt (Zaharchuk et al., 2018; Zhao et al., 2017). Unter Dropout versteht man eine Technik während des Trainingsprozesses, bei der die Ausgabe zufällig ausgewählter Neuronen verworfen wird, um die Generalisierung zu verbessern (Srivastava et al., 2014).

Eine weitere mögliche Lösung ist die Regularisierung, dazu gehören eine Reihe von Methoden, die die Komplexität des Modells verringert. Dies geschieht durch die Verringerung der Werte der Gewichte (Zaharchuk et al., 2018).

Trainingszeit

Die Verringerung der Trainingszeit und eine schnellere Konvergenz wird in zahlreichen Studien untersucht. Konvergenz beschreibt den Zustand, in dem zusätzliches Training das Modell nicht weiter verbessert. Eine möglicher Lösungsansatz ist die Anwendung von Pooling-Schichten, um die Dimensionalität der Parameter zu reduzieren (Dou et al., 2017). Moderne pooling-basierte Lösungen verwenden die Faltung mit Stride (Simonyan & Zisserman, 2014) die die Bilddaten komprimieren und somit das Netzwerk schlanker machen. Unter Stride versteht man die Anzahl der Schritte, die jeweils zurückgelegt werden, also die Schrittbreite (Simonyan & Zisserman, 2014). Ein Nachteil von pooling-basierten Lösungen ist der Informationsverlust (Hesamian et al., 2019).

Ein weiterer Lösungsansatz ist die Batch-Normalisierung: Durch die Standardisierung der Eingaben in eine Schicht für kleine Stapel wird der Trainingsprozess stabilisiert und es werden weniger Epochen benötigt (Baumgartner et al., 2018; Çiçek et al., 2016; Ioffe &

Szegedy, 2015; Kawahara et al., 2016). Für die Batch-Normalisierung wurden keine negativen Auswirkungen auf die Leistung berichtet, daher ist sie der beliebteste Ansatz zur Reduzierung der Trainingszeit (Hesamian et al., 2019).

Komplexität in der Radiologie

Der Unterschied in der Radiologie im Vergleich zu anderen Bilderkennungsanwendungen von Deep Learning Algorithmen besteht darin, dass eine CT- oder MRT-Untersuchung aus tausenden Bildern bestehen kann, nicht nur aus einem einzigen Bild. Daher sind die erforderlichen Algorithmen zwangsläufig wesentlich komplexer (McBee et al., 2018).

Darüber hinaus ist anzumerken, dass viele andere Anwendungen, wie z. B. die Gesichtserkennung, mit einem relativ homogenen Datensatz arbeiten, während Bilder in der Radiologie je nach Patientenfaktoren und Pathologien stark variieren können, was die Komplexität weiter erhöht (McBee et al., 2018). Eine der größten Herausforderungen bei der medizinischen Bildsegmentierung ist die Heterogenität menschlicher Organe und Läsionen. Das Erscheinungsbild von Organen und Läsionen kann in Form, Größe und Lage stark variieren, abhängig von der medizinischen Vorgeschichte des Patienten und den Eingriffen, denen der Patient vor der Aufnahme des Bildes unterzogen wurde. Dies stellt eine Herausforderung für Deep Learning Netzwerke dar (Hesamian et al., 2019). Es hat sich jedoch gezeigt, dass dieses Problem durch tiefere Netzwerke mit mehr Schichten gelöst werden kann (Hesamian et al., 2019; Yu et al., 2017).

1.1.5 Historische Entwicklung

KI hat das Potential das Gesundheitswesen und die Medizin zu revolutionieren. Vor allem durch den Zugang zu größeren Bildgebungsdatensätzen konnten in den letzten Jahren

große Erfolge in der Entwicklung von Algorithmen verzeichnet werden (Mazurowski et al., 2019). Für komplexere Fragestellungen und Algorithmen wird immer leistungsstärkere Computerhardware benötigt. Diese wurde in den letzten Jahren immer günstiger und gleichzeitig leistungsstärker. Wie bereits beschrieben hat die Einführung und Nutzung von Grafikprozessoren (GPUs) die Rechenzeit drastisch reduziert, wodurch die Entwicklung in diesem Bereich weiter vorangetrieben wurde (Mazurowski et al., 2019). Die Verfügbarkeit von Open-Source-Software-Frameworks erleichtert den Fortschritt ebenfalls erheblich. Allerdings müssten "fertige" Systeme angeboten werden, die in bestehende Arbeitsabläufe integriert werden können, um diese Methoden routinemäßig in die klinische Praxis einzubinden. Je vertrauter Wissenschaftler*innen und Radiolog*innen selbst mit Deep Learning Algorithmen werden, desto einfacher werden komplexe und spezielle Probleme mit Deep Learning Algorithmen gelöst werden können (Zaharchuk et al., 2018).

Derzeit kann ein klarer Aufwärtstrend bezüglich des Einsatzes von KI verzeichnet werden. Die Fülle an möglichen Anwendungen für Deep Learning Algorithmen wird wahrscheinlich auch in Zukunft zu einem verstärkten Einsatz dieser Technologie führen. Ganz entscheidend für den Erfolg von Deep Learning Algorithmen ist die Größe der verfügbaren Datenmenge. Je mehr beschriftete Datensätze zur Verfügung stehen, desto leistungsfähiger werden Deep Learning Ansätze sein. Zentralisierte Datensammlungen und -freigaben wie das Cancer Imaging Archive (*The Cancer Imaging Archive*, 2023) sind in dieser Hinsicht vielversprechend. Das exponentielle Wachstum der verfügbaren Datensätze, das in den letzten Jahren verzeichnet wurde, lässt sich zum Teil durch die größere Verfügbarkeit von multiplanaren, multikontrastreichen und mehrphasigen MRTs

erklären. Darüber hinaus spielt die Bildgebung eine immer wichtigere Rolle bei der Diagnose in fast allen Bereichen der Medizin, wodurch automatisch mehr Daten erzeugt werden (Zaharchuk et al., 2018).

Es hat sich schnell herauskristallisiert, dass eine der größten Stärken von Deep Learning Algorithmen in der Medizin im Bereich der Bildanalyse liegt. Die Radiologie ist von Natur aus ein datengesteuerter Bereich, daher bietet sich der Einsatz von KI hier hervorragend an. Nach der fast vollständigen Umstellung von analogen Optionen auf zentrale digitale Bildarchivierungs- und Bildbetrachtungssysteme (PACS) sind im Gegensatz zu anderen Bereichen der Medizin heute fast alle erzeugten und verwendeten Daten, sowie die von Radiologen erstellten Befunde primär digital (Mazurowski et al., 2019).

Deep Learning Modelle sind eine der am weitesten fortgeschrittenen Ansätze des maschinellen Lernens. Diese konnten von Beginn an großartige Erfolgsquoten für Mustererkennungsaufgaben vorweisen. Schon frühe Studien, in denen Deep Learning Algorithmen für die Erkennung oder Klassifizierung von Läsionen angewendet wurde, berichteten von überdurchschnittlichen Leistungen im Vergleich zu konventionellen Techniken und teilweise sogar von besserer Leistung als Radiolog*innen bei bestimmten Aufgaben (Chan et al., 2020). Ein weiterer interessanter Aspekt ist der unvoreingenommene Blickwinkel, mit dem Deep Learning Netzwerke Merkmale begutachten und somit möglicherweise für uns neue Prädiktoren definieren. Beispielsweise hat ein Machine Learning Algorithmus, der trainiert wurde, um die Prognose von Burstkrebspatienten zu bestimmen, nicht nur auf die histologischen Aspekte der Tumorzellen, auf die auch Patholog*innen achten würden, sondern auch auf das umliegende Stroma geachtet (Beck et al., 2011; Olczak et al., 2017)

1.1.6 Aktueller Wissensstand: Deep Learning in der Radiologie

In den letzten Jahren ist die Zahl der veröffentlichten wissenschaftlichen Arbeiten, in denen vielversprechende Ergebnisse publiziert wurden und die den Erfolg von Deep Learning Modellen in der Radiologie belegen, rasant angestiegen. Um die vielfältigen Einsatzmöglichkeiten zu veranschaulichen, soll im Folgenden ein Überblick über die publizierten Arbeiten gegeben werden.

Ein CNN-Modell wurde entwickelt, um klinisch signifikante Befunde in CT-Scans des Kopfes zu identifizieren. Das Modell wurde mit 3,5 Millionen CT-Bilder aus über 24.000 Untersuchungen trainiert. Es wurden 30 phänomenologische Merkmale definiert, die erkannt werden sollten. Dieses Modell erreichte eine geringere klinisch signifikante Fehlerrate (0,0367 %) als US-amerikanische Fachärzt*innen für Radiologie (0,82 %) (Merkow et al., 2017).

Bei einem weiteren Deep Learning Ansatz für die automatisierte Erkennung von Hirnblutungen auf der Basis von CT-Datensätzen wurde versucht die Vorgehensweise der Radiolog*innen nachzuahmen. Der Algorithmus durchsuchte die einzelnen 2D-Querschnitte nach Blutungsregionen und zog die benachbarten Schichten mit in die Analyse ein, um dann wiederum die Vorhersage auf Schichtebene zu machen, um eine Diagnose zu stellen. Beim Vergleich von 77 Schädel-CTs des Modells mit drei erfahrenen Radiolog*innen konnte das Modell mit einer Vorhersagegenauigkeit von 81,82% überzeugen, die vergleichbar mit der von Radiolog*innen ist. Außerdem hatte das Deep Learning Modell eine höhere Sensitivität als zwei der drei Radiolog*innen (Grewal et al., 2018).

Ein weiteres Modell wurde trainiert, um mithilfe von T1-gewichteten MRT Bildern Knorpelknochentumore als niedrig- bis hochgradig maligne zu klassifizieren. Es konnte hier kein signifikanter Unterschied in der Leistung zwischen der Radiolog*in und dem Klassifikator des maschinellen Lernens (P = 0,453) festgestellt werden (Gitto et al., 2020). Dieses Modell könnte sich als wertvolle Hilfe bei der präoperativen Tumorcharakterisierung erweisen.

Des Weiteren konnte ein CNN Deep Learning Modell das Knochenalter anhand von konventionellen Röntgenbildern der Hand bestimmen. Die Genauigkeit war mit einer erfahrenen Radiolog*in zu vergleichen (Larson et al., 2018).

Ein anderes Deep Learning Netzwerk wurde trainiert, um Hüftfrakturen auf Röntgenbildern zu erkennen. Auch dieses Deep Learning Netzwerk erreichte Ergebnisse, die mit denen von Radiolog*innen zu vergleichen waren. Hier ist jedoch anzuführen, dass die Radiolog*innen Zugriff auf weitere Scans der Patienten hatten, das DL-Netzwerk nicht, was zu einer eingeschränkten Vergleichbarkeit und Aussagekraft führt (Gale et al., 2017).

Im Jahre 2020 wurde von von Schacky et al. ein Deep Learning Modell vorgestellt, das mehrere Klassifizierungsaufgaben gleichzeitig lösen konnte. Dieses Modell wurde anhand von 15364 Hüftgelenken und fünf Hüftarthrose-Merkmalen (femorale Osteophyten, acetabuläre Osteophyten, Gelenkspaltverengung, subchondrale Sklerose und subchondrale Zyste) pro Röntgenbild trainiert. Dieses Modell konnte die Merkmale so zuverlässig wie eine Oberärzt*in der Radiologie bewerten (von Schacky et al., 2020). Je nach Merkmal variierte die Genauigkeit des Modells: 89 % für femorale Osteophyten,

76 % für acetabuläre Osteophyten, 83 % für Gelenkspaltverengung, 96 % für subchondrale Sklerose und 97 % für subchondrale Zysten (von Schacky et al., 2020).

Jamaludin et al. stellte ein weiteres Modell vor, das mithilfe von MRT-Scans trainiert wurde. Es sollte eine Automatisierung der Klassifizierung von lumbalen Bandscheiben und Wirbelkörpern aus MRT-Scans der Lendenwirbelsäule stattfinden, was erfolgreich war. Das Modell konnte Vorhersagen für mehrere pathologische Einstufungen treffen, die mit Einschätzung von Radiolog*innen übereinstimmten (Jamaludin et al., 2017).

Ferner konnte ein zur Klassifizierung von Weichteilsarkomen trainiertes Deep Learning Netzwerk erfolgreich den Differenzierungsgrad von Tumoren auf MRT T1-gewichteten und T2-gewichteten Sequenzen beurteilen. Ein nichtinvasives Tumorgrading ohne Biopsie ist damit also möglich. Dies erspart den Patient*innen die invasive Biopsie, die wie jeder invasive Eingriff mit Risiken und Schmerzen verbunden ist. Interessanterweise konnte das T2 basierte Modell außerdem Patient*innen mit hohem Sterberisiko nach Therapie statistisch signifikant identifizieren (Navarro et al., 2021).

Ein weiterer Algorithmus, der für die Klassifizierung von primären Knochentumoren entwickelt wurde, wurde mithilfe eines multizentrischem Datensatzes an präoperativen konventionellen Röntgenbildern trainiert. Dieser Algorithmus konnte primäre Knochentumore besser als Juniorradiolog*innen und genauso gut wie Fachärzt*innen klassifizieren (He et al., 2020).

Ein Fusionsmodell aus Deep Learning Modellen und maschinellem Lernen zur Klassifizierung von gutartigen, bösartigen und intermediären Knochentumoren wurde im Jahre 2022 vorgestellt: Dieses Modell wurde mithilfe von klinischen Patientenmerkmalen

und konventionellen Röntgenbildern trainiert. Das Niveau des Fusionsmodells war mit dem Niveau von erfahrenen Radiolog*innen vergleichbar (Liu et al., 2022).

Auch in Hinblick auf Wirbelkörperfrakturen gibt es mehrere Studien, die in den letzten Jahren veröffentlicht wurden. Dass Deep Learning Algorithmen Wirbelkörperfrakturen erkennen können, wurde schon mehrfach bewiesen:

Ein CNN wurde zur Diagnose von Wirbelkörperfrakturen trainiert. Das Modell wurde mit thorakolumbalen Röntgenbildern von 300 Patient*innen, jeweils 150 Patient*innen mit und 150 Patient*innen ohne Wirbelkörperfrakturen trainiert. Das CNN erreichte Genauigkeits-, Sensitivitäts- und Spezifitätsraten von 86 %, 85 % bzw. 87 %, die denen der orthopädischen Chirurg*innen nicht unterlegen waren. Dieses CNN kann bei der frühzeitigen Diagnose von Wirbelkörperfrakturen helfen (Murata et al., 2020).

Ein weiteres auf KI basierendes Modell wurde ebenfalls zur Diagnose von Wirbelkörperkompressionsfakturen entwickelt. Diese multizentrische, pro- und retrospektive Studie basiert auf einem Modell, das anhand von 1904 Patient*innen trainiert wurde. Das Modell erkannte Wirbelfrakturen auf Röntgenbildern in sagittaler Ansicht mit hoher Genauigkeit, Sensitivität und Spezifität, insbesondere osteoporotische Lendenwirbelfrakturen (Xu et al., 2023).

Weiterhin wurde ein CNN vorgestellt, das osteoporotische Wirbelfrakturen in CT-Untersuchungen erkennen kann. Das CNN wurde mithilfe von 1.432 CT-Scans mit 10.546 zweidimensionalen Bildern in sagittaler Ansicht trainiert und bewertet. Die Genauigkeit wurde mit 89 % angegeben und der F1-Score mit 91 %. Die Ergebnisse dieses Systems entsprachen der Leistung von Radiolog*innen unter realen klinischen

Bedingungen. Dadurch ist dieses CNN geeignet, die Diagnose von osteoporotischen Wirbelfrakturen im klinischen Alltag zu unterstützen und zu verbessern, indem es CT-Untersuchungen vorab überprüft und potentielle Wirbelkörperfrakturen kennzeichnet (Tomita et al., 2018).

Nicht nur Wirbelkörperfrakturen konnten als solche erkannt werden, es wurde auch bereits ein Modell vorgestellt, das maligne Läsionen in Wirbelkörpern erkennen kann: Dieses Modell konnte erfolgreich metastatische von nicht-metastatischen Wirbelkörpern in MRT-Scans unterscheiden. Es wurden sowohl T1-gewichtete als auch T2-gewichtete Sequenzen untersucht. Die wichtigsten Prädiktoren basierten auf der T2-gewichteten Sequenz und waren morphologische und texturelle Merkmale (Filograna et al., 2019).

Dies konnte auch anhand von konventionellen Röntgenbildern für die Differenzierung von gutartigen und bösartigen Knochenläsionen gezeigt werden: Ein maschinelles Lernmodell konnte erfolgreich zwischen gutartigen und bösartigen Knochenläsionen unterscheiden. Das ANN wurde einerseits mithilfe der radiologischen Merkmale und andererseits mithilfe von demografischen Informationen trainiert. Durch die Inkludierung der demographischen Information konnte die bestmögliche Leistung erreicht werden. Dies unterstreicht die Bedeutung der Entwicklung umfassender Modelle. Dieses Modell war bezüglich der Genauigkeit Assistenzärzt*innen überlegen, aber auf muskuloskelettale Tumore spezialisierten Radiolog*innen unterlegen (von Schacky et al., 2022).

Zusammenfassend ist also zu sagen, dass viele Deep Learning Algorithmen bestimmte Merkmale auf ähnlicher Ebene wie Assistenzärzt*innen, teilweise sogar wie

Fachärzt*innen bewerten können. Es ist jedoch zu beachten, dass der Zeitaufwand den Algorithmus zu trainieren in keiner Weise mit der jahrelangen Ausbildung bis zum Facharzt eines Radiologen zu vergleichen ist. Weiterhin ist die ungeklärte Frage nach der Zuverlässigkeit zwischen verschiedenen Beobachter*innen, die Interobserver-Reliabilität, nicht zu unterschätzen (Sayed-Noor et al., 2011; Shehovych et al., 2016). Die Objektivität und Schnelligkeit von Deep Learning Modellen bringen wertvolle Chancen mit sich und sind aus klinischer Sicht hochinteressant (Jamaludin et al., 2017; Olczak et al., 2017).

1.1.7 KI in der Radiologie: Herausforderungen und Lösungsansätze

Die Einführung von Deep Learning Algorithmen in der Radiologie wird unabhängig von den genauen Details einige Herausforderungen mit sich bringen.

Die größte Herausforderung ist im Moment der technische Aspekt. Obwohl Deep Learning bei vielen anderen bildbezogenen Aufgaben äußerst erfolgreich ist und vielversprechende Ergebnisse geliefert hat, sind die Algorithmen in der Radiologie derzeit nicht in der Lage, Radiolog*innen vollständig zu ersetzen. Nach aktuellem Stand der Studien können Algorithmen heute zwar einzelne eng definierte Fragestellungen mit der Expertise eines menschlichen Experten klären, aber diese Ergebnisse lassen sich nur auf einen sehr kleinen Teil der Fragestellungen anwenden, mit denen sich Radiologen beschäftigen müssen. Nach den Fortschritten der letzten Jahre zu urteilen, wird sich dies in den kommenden Jahren weiter rasant verbessern (Gale et al., 2017; Grewal et al., 2018; Jamaludin et al., 2017; Kooi et al., 2017; Larson et al., 2018; Mazurowski et al., 2019; Merkow et al., 2017; Olczak et al., 2017; Rajpurkar et al., 2017).

Eine weitere Herausforderung ist das Vertrauensverhältnis zwischen Patient*innen und ihren behandelnden Ärzt*innen. Es wird die Frage nach Akzeptanz seitens der Patient*innen gestellt, wenn der Prozess der Bildinterpretation ohne die Beteiligung einer Radiolog*in erfolgt. Es wurde ein signifikant höheres Vertrauen in die Interpretation und Diagnose durch eine Radiolog*in als in KI-gestützte Interpretationen gezeigt. In einer Studie gaben 95,4 % der Befragten an, dass sie bei Diskrepanzen die klinische Interpretation der Radiolog*in der KI-Interpretation vorziehen würden (York et al., 2020). Auch ethische und rechtliche Fragen müssen geklärt werden. Beispielsweise ist die Frage, wer für etwaige Fehldiagnosen, die theoretisch zu nicht angebrachten oder falschen Therapien führen, verantwortlich ist, noch nicht abschließend geklärt.

Als weitere Herausforderung kann man den Wissensstand der zukünftigen Radiolog*innen anführen. Falls Deep Learning Modelle im klinischen Alltag mehr und mehr eingesetzt werden und selbstständig befunden können und dürfen, wird damit automatisch ein reduziertes Arbeitspensum der Radiolog*innen mit einhergehen. Es wirft die Frage auf, inwieweit das Wissen der zukünftigen Radiolog*innen dadurch eingeschränkt wird und ob sich Mediziner*innen in Zukunft verleiten lassen, sich zu sehr auf KI zu verlassen und somit zu abhängig von dieser neuen Technologie werden.

Derzeit wird von Deep Learning Algorithmen in der Radiologie normalerweise nur ein sehr kleiner Ausschnitt betrachtet. Die klinische Symptomatik und der Kontext wird nicht mit in den Entscheidungsprozess mit einbezogen. Es ist wichtig in Zukunft einen Weg zu finden, diese Information mit einzubeziehen, sodass komplexe medizinischen Fragestellungen richtig beurteilt werden können und keine Fehldiagnose gestellt wird. Dies könnte mithilfe von Fusionsmodellen gelingen.

Ein weiterer wichtiger ungeklärter Punkt ist, dass Deep Learning Modelle für uns derzeit noch mit großem Unverständnis verbunden sind. Was in dieser „Black Box“ bei undurchschaubaren Modellen geschieht, ist weiterhin nicht vollständig geklärt. Als Antwort auf diese Fragestellung kam es zu einem großen Interesse an erklärbarer KI (XAI). Hierunter versteht man sowohl inhärent erklärbare Techniken als auch Versuche die Entscheidungsfindung von Black-Box-KI-Systeme für Menschen zu veranschaulichen und zu erklären. Das Problem dieser Ansätze ist allerdings die Gefahr, nur eine scheinbare Transparenz zu schaffen, ohne wirklich echtes Verständnis zu erlangen. Dadurch würden sich menschliche Entscheidungsträger in falscher Sicherheit wiegen (Cabitza et al., 2022).

Finlayson et al. macht auf eine ganze Reihe potenziell gefährlicher Anfälligkeiten von KI aufmerksam. Er beschreibt „adversarial attacks“. Darunter versteht man Manipulationen, die explizit darauf ausgelegt sind Algorithmen zu täuschen und dazu führen, dass Algorithmen zu einem falschen Ergebnis kommen (Finlayson et al., 2019). Es konnten diese Art von Angriffen für praktisch jede Art von maschinellem Lernmodell, das jemals untersucht wurde, gefunden werden. Dazu gehören eine breite Palette an Datentypen, einschließlich Bildern, Audio, Text und anderen Eingaben (Biggio & Roli, 2018). Es ist wichtig anzumerken, dass dies keine Ungenauigkeit oder Unzuverlässigkeit von KI, sondern eine Anfälligkeit für Manipulationen, die explizit darauf ausgelegt sind sie zu täuschen, widerspiegeln. Diese Techniken könnten beispielsweise eingesetzt werden, um bei Abrechnungscodes die bestmögliche Codekombination herauszufinden, die die Erstattung maximiert oder das Risiko einer Ablehnung des Antrags minimiert (Biggio & Roli, 2018). Das Problem geht in der Medizin jedoch weit über potentiellen

Versicherungsbetrug hinaus und erfasst ein breites Spektrum an Handlungsmotiven. Beispielsweise könnten auch Angriffe durch den Wunsch nach einer hochqualitativen Versorgung motiviert sein. Es wurde bereits versucht Algorithmen zu entwickeln, die nicht anfällig für diese Art von gegnerischen Angriffen sind. Diese Algorithmen wurden mit gegnerischen Beispielen trainiert oder die Daten alternativ verarbeitet, um potenzielle Manipulationen abzuwehren. Eine noch ungeklärte, aber wesentliche Frage ist, wann und in welcher Art und Weise man hier eingreifen und dagegenwirken sollte (Finlayson et al., 2019).

1.1.8 Ausblick in die Zukunft

Die Nachfrage nach bildgebenden Verfahren wird höchstwahrscheinlich in den kommenden Jahren weiter steigen. Gleichzeitig wird ein Mangel an ausgebildeten Radiolog*innen erwartet. Dies ist auf die große Zahl von Radiolog*innen zurückzuführen, die in naher Zukunft das Rentenalter erreichen werden, ohne dass es ausreichend Nachwuchskräfte gibt, die diese Positionen füllen könnten. KI hat das Potenzial, diese Probleme zu lösen (D. H. Kim & MacKinnon, 2018).

KI hat in den letzten Jahren rasante Fortschritte verzeichnet. Geht man davon aus, dass die Entwicklung in ähnlichem Tempo wie in den letzten Jahren voranschreiten wird, kann man mit einer großen Weiterentwicklung rechnen (Zaharchuk et al., 2018). Insbesondere Deep Learning Modelle werden im Vergleich zu anderen medizinischen Fachgebieten wahrscheinlich zuerst die Radiologie betreffen und die tägliche Arbeit von Radiolog*innen tiefgreifend verändern (Zaharchuk et al., 2018).

Deep Learning hat sich in den letzten Jahren zu einem außergewöhnlich leistungsfähigen Werkzeug für die Bildverarbeitung entwickelt. Trotz der sich abzeichnenden Entwicklung

spielen Algorithmen derzeit im Alltag von Kliniken und radiologischen Praxen in Deutschland eine untergeordnete Rolle und werden nicht routinemäßig eingesetzt. Aktuelle Erfolge lassen uns jedoch auf Algorithmen hoffen, die in die tägliche Routine klinisch tätiger Radiolog*innen integriert werden können (Chan et al., 2020). Eine wichtige Frage ist, wie diese Integration in den klinischen Alltag gelingen kann. Weitere Forschungen zur Beziehung zwischen Radiolog*innen und KI sind hierzu erforderlich. Beispielsweise sollte die Frage geklärt werden, wie Radiolog*innen den Umgang mit KI-Tools erlernen und die Ergebnisse der KI-Tools richtig interpretieren können.

Generell besteht sowohl unter wissenschaftlich als auch klinisch tätigen Ärzt*innen ein Konsens darüber, dass KI, insbesondere Deep Learning Algorithmen, mittel- bis langfristig eine wichtige Rolle in der Bildgebung spielen werden (Recht & Bryan, 2017). Weiterhin besteht Konsens darüber, dass die Einbeziehung von KI in die Radiologie die diagnostische Genauigkeit verbessern würde (Recht & Bryan, 2017). Solche Instrumente müssen jedoch sorgfältig untersucht werden, bevor sie in den klinischen Alltag integriert werden (Kalmet et al., 2020).

Eine mögliche Option wäre, dass Algorithmen banale Aufgaben übernehmen, wie beispielsweise redundante Mustererkennungsaufgaben, während sich Radiolog*innen auf kognitiv anspruchsvollere Fragestellungen konzentrieren (Jha & Topol, 2016).

Eine weitere Option wäre, dass KI Radiolog*innen unterstützt und ihnen zuarbeitet. Hier wäre es hilfreich, wenn Radiolog*innen über ein grundlegendes Verständnis von KI und KI-basierten Werkzeugen verfügen. Diese Werkzeuge würden die Arbeit der Radiolog*innen nicht ersetzen, und ihre Rolle wäre nicht notwendigerweise auf die

Interpretation von Befunden beschränkt. KI-Tools könnten als ergänzende Instrumente eingesetzt werden, um die Entscheidungen der Radiolog*innen zu bestätigen und zu validieren (Liew, 2018). Weiterhin ist es denkbar, dass KI die Aufmerksamkeit der Radiolog*innen auf suspekte Bildbereiche lenkt (Jamaludin et al., 2017). Beispielsweise könnte durch die Vorprüfung von MRT- und CT-Bildern durch einen DL-Algorithmus und die Kennzeichnung kritischer Befunde die Geschwindigkeit der Befundung der dringendsten Fälle erhöht und die diagnostische Genauigkeit von unklaren Befunden verbessert werden (Hirschmann et al., 2019). Weiterhin könnte dies ermüdungsbedingten Diagnosefehlern im Nachtdienst oder bei der Befundung großer Mengen von z. B. postoperativen Röntgenbildern oder Staging-Untersuchungen entgegenwirken. Die Hoffnung ist, dass sich dadurch die Qualität der Diagnosen und Interpretationen verbessert (Hirschmann et al., 2019).

Eine dritte Option wäre, dass Deep Learning Algorithmen und Radiolog*innen gleichberechtigt zusammenarbeiten und so bessere Leistungen erbacht werden.

Eine vierte Möglichkeit wäre, dass Radiolog*innen vollständig ersetzt werden, insbesondere wenn man ihre Funktion als Bildinterpret*innen betrachtet. Algorithmen könnten auch als Unterstützung bei der Diagnosefindung in entlegenen Gebieten ohne radiologische Spezialisten oder in teleradiologischen Diensten dienen, wenn keine Radiolog*in vor Ort sein kann. Auf diese Weise könnte eine flächendeckende gute Versorgungsqualität gewährleistet werden (Hirschmann et al., 2019). Von vielen wird als mittel- oder langfristiges Ziel End-to-End-CNNs gesehen. Darunter versteht man Folgendes: Scans werden als Input verwendet und der Output ist ein vollständiger

Bericht, der dem Bericht einer menschlichen Radiolog*in gleicht (Chen et al., 2022; Guo et al., 2021).

Obwohl derzeit KI in der Radiologie in erster Linie auf die Erkennung, Diagnose und Charakterisierung von Krankheiten ausgerichtet ist, bietet sie auch eine Möglichkeit zur Erleichterung der Arbeitsabläufe. Techniken des maschinellen Lernens könnten auf nicht bildgebende radiologische Aufgaben angewendet werden. Beispielsweise könnten Deep Learning Algorithmen in der Radiologie auf Textdaten angewendet werden, um das Protokoll zu bestimmen. Die Entscheidung, welche Sequenzen benötigt werden, würde dadurch automatisiert werden. Die Bestimmung des MRT-Protokolls ist ein wesentlicher Prozess im Arbeitsablauf, und die Auswahl des am besten geeigneten Protokolls ist entscheidend für genaue radiologische Interpretationen und eindeutige radiologische Diagnosen. Diese Vorarbeit ist sehr zeitaufwändig und führt mitunter zu einer erheblichen Arbeitsbelastung (Lee, 2018). In Bezug auf die Rückrufe von Patient*innen zur Wiederholung einer Untersuchung sind etwa 20 % auf Protokollfehler in der Ambulanz zurückzuführen (Gyftopoulos et al., 2016). Deep Learning könnte diese Aufgabe übernehmen und zur Beseitigung von Fehlerquellen führen, indem es zeitnahe und hochpräzise Protokollfestlegungen liefert, die nur eine schnelle Bestätigung erfordern. Um die Effizienz in der Klinik oder radiologischen Praxis zu verbessern wäre die Einführung eines Deep Learning CNNs zur Protokollbestimmung mit einem auf Kurztext-Klassifikation basierenden Klassifikator zu überlegen. In einer Studie, in der genau diese Fragestellung untersucht wurde, wurde festgestellt, dass alle MRT-Protokolle bei den MRTs von Beckenknochen, Oberarm, Handgelenk und Unterschenkel korrekt

ausgewählt wurden, die Protokolle also mit denen, die von Radiolog*innen ausgewählt wurden, übereinstimmten (Lee, 2018). CNNs können also zur Festlegung von MRT-Protokollen eingesetzt werden. Die Ausweitung von CNN-basierten Anwendungen auf andere radiologische Aufgabenfelder neben der Bildinterpretation bedeutet eine große Zeitersparnis für Radiolog*innen (Lee, 2018) und gleichzeitig mehr Zeit für die Patient*innenversorgung (McBee et al., 2018).

1.2 Osteoporotische und pathologische Wirbelkörperfrakturen

1.2.1 Differentialdiagnosen und Prävalenz

Vertebrale Kompressionsfrakturen (VCFs) können verschiedenste Ursachen haben. Dazu gehören Traumata, genetische oder metabolische Ursachen, Osteoporose oder eine maligne Grunderkrankung (Geith et al., 2015). Abgesehen von seltenen metabolischen und genetischen Erkrankungen muss bei atraumatischen Wirbelkörperfrakturen grundsätzlich zwischen gutartigen osteoporotischen und pathologischen Wirbelbrüchen unterschieden werden. Diese atraumatischen Wirbelkörperkompressionsfrakturen in der Brust- und Lendenwirbelsäule sind ein bekanntes klinisches Problem, insbesondere bei älteren Patient*innen (Geith et al., 2015).

Die Prävalenz von Osteoporose hat in Deutschland in den letzten Jahren aufgrund des demografischen Wandels stetig zugenommen. Mit der vermehrten Alterung der Gesellschaft geht auch eine steigende Zahl an osteoporotischen Frakturen einher. Osteoporose stellt die häufigste Ursache für Wirbelkörperfrakturen bei älteren Patient*innen dar (Geith et al., 2015). Wirbelkörperfrakturen sind die häufigste Form von

osteoporotischen Frakturen und gehen mit einer erheblichen Morbidität (Ettinger et al., 1992; Lentle et al., 2019; Nevitt, 1998) und erhöhten Mortalität einher (Cooper, Atkinson, et al., 1993; Fink et al., 2003; Kendler et al., 2016; Lentle et al., 2019). Sie stellen außerdem einen wichtigen Risikofaktor für weitere Wirbelfrakturen dar (Lentle et al., 2019).

Wie groß der Anteil der Bevölkerung ist, die unter osteoporotischen Wirbelkörperfrakturen leidet, wurde in zahlreichen Studien untersucht. Grundsätzlich sind Männer statistisch gesehen seltener betroffen, was durch hormonelle Unterschiede erklärt werden kann (J. L. Melton, 1997).

Die kanadische Multicentre Osteoporosis Study berichtete, dass 21,5 % der Männer und 23,5 % der Frauen über 50 Jahren mindestens eine Wirbelkompressionsdeformität aufweisen, (Jackson et al., 2000) während die norwegische Tromso-Studie ergab, dass 20,3 % der Männer und 19,2 % der Frauen über 70 Jahren mindestens eine Wirbelfraktur hatten (Waterloo et al., 2012). In einer Studie in Rochester, Missouri, wurden bei einem Viertel der Frauen im Alter von über 50 Jahren eine oder mehrere Wirbelkörperdeformierungen diagnostiziert (L. J. Melton et al., 1993). In der Study of Osteoporotic Fractures erlitten 18% der postmenopausalen Frauen über 65 Jahre in einem 15-jährigen Follow-up eine Wirbelfraktur (Cauley et al., 2007). Zwischen 10 % und 28 % der Wirbelfrakturen wurden bei postmenopausalen Frauen mit einem T-Score der Knochenmineraldichte > -2,5 festgestellt (Greenspan et al., 2001; Schousboe et al., 2006). Obwohl unterschiedliche Definitionen von Wirbelbrüchen den Vergleich von Daten aus verschiedenen Untersuchungen erschweren, hat mindestens einer von fünf Männern und Frauen im Alter von über 50 Jahren eine oder mehrere Wirbelbrüche.(Kendler et al.,

2016). In einer weiteren Studie wird die Prävalenz osteoporotischer Wirbelkörperfrakturen bei postmenopausalen Frauen mit etwa 25% angegeben (J. L. Melton, 1997).

Trotz der hohen Prävalenz von Wirbelfrakturen bleiben mehr als zwei Drittel der Wirbelfrakturen unerkannt (Cooper, O'Neill, et al., 1993; Fink et al., 2005). Die Erkennung von Wirbelfrakturen in Bildgebungen, die zu anderen Zwecken als der Untersuchung auf Wirbelfrakturen in einem Krankenhaus erstellt wurden, ist generell nicht zufriedenstellend (Bartalena et al., 2009; Cataldi et al., 2008; Fernández et al., 2012; Kendler et al., 2016; N. Kim et al., 2004; Majumdar et al., 2005; Mui et al., 2003; Obaid et al., 2008; Williams et al., 2009; Woo et al., 2008).

Gleichzeitig nimmt aufgrund der Alterung der Gesellschaft und der fortschreitenden medizinischen Möglichkeiten in der Krebstherapie die Prävalenz metastasierender Malignome ebenfalls zu. Metastasen verschiedener primärer Krebsarten treten am häufigsten in der Leber, der Lunge und an dritter Stelle im Skelettsystem auf (Filograna et al., 2019). Insgesamt finden sich 30% bis 55% der Metastasen im Skelettsystem in der Wirbelsäule (Aebi, 2003; Jung et al., 2003; W. H. Kim et al., 1995). Brust-, Lungen- und Prostatakrebs sind die häufigsten Krebsarten, die in die Wirbelsäule metastasieren. Diese machen zusammen 60% aller Wirbelsäulenmetastasen aus (Aebi, 2003). Abgesehen von Metastasen gehören zu den Ursachen pathologischer Wirbelfrakturen die folgenden malignen Erkrankungen: Multiples Myelom, primäre bösartige Knochentumore, primäre oder sekundäre Lymphome und diffuse Knochenmarkserkrankungen, wie beispielsweise Leukämien und myeloproliferative Erkrankungen (Karchevsky et al., 2008).

1.2.2 Diagnostik

Pathologische Kompressionsfrakturen und deren Unterscheidung von gutartigen osteoporotischen Kompressionsfrakturen haben Auswirkungen auf das Staging, die Erstellung eines geeigneten Behandlungsplans und die Prognose von Patient*innen mit maligner Grunderkrankung (Jung et al., 2003). Genauso wichtig ist die korrekte Diagnose osteoporotischer Wirbelfrakturen, da sie ein zuverlässiger Prädiktor für künftige Frakturen sind und zu einer verkürzten Lebenserwartung führen (Kendler et al., 2016).

Wirbelkörperfrakturen können sich klinisch in unterschiedlichster Weise zeigen. Die Frakturen können zu akuten und chronischen Schmerzen führen (Fink et al., 2003), die mit einer Beeinträchtigung der Lebensqualität einhergehen (Kendler et al., 2016) und sich mit akuten, schmerzhaften Wirbelverformungen präsentieren (Torres & Hammond, 2016). Diese Schmerzen können über mehrere Monate anhalten (Fink et al., 2003). Eine akute Wirbelfraktur kann außerdem von Muskelkrämpfen begleitet sein (Fink et al., 2003). Weiterhin können pathologische Frakturen der Wirbelsäule zu einem insgesamt geschwächten Allgemeinzustand aufgrund starker Schmerzen und möglicher motorischer Schwäche führen (Cho et al., 2015).

Ein Großteil der pathologischen und osteoporotischen Wirbelkörperfrakturen bleibt jedoch unbemerkt und wird nicht klinisch auffällig (Cho et al., 2015; Kendler et al., 2016). Dies ist einer der Gründe, warum sie grundsätzlich unterdiagnostiziert werden (Tomita et al., 2018).

Beide Entitäten, sowohl osteoporotische als auch pathologische Wirbelkörperfrakturen treten in der Regel bei älteren Erwachsenen auf, meistens im thorakalen und lumbalen

Bereich der Wirbelsäule, können koexistieren und können bei normaler physiologischer Belastung auftreten (Torres & Hammond, 2016).

Weiterhin ist anzuführen, dass ein neuer Wirbelbruch bei einer Patient*in mit einer Krebserkrankung in der Vorgeschichte zwar vermuten lässt, dass es sich um eine Metastase handelt, doch ist dies nicht zwangsweise der Fall. Eine Studie in der 659 Krebspatienten postmortal einer Autopsie unterzogen wurden, hat gezeigt, dass bei einem Drittel der Krebspatienten ein frakturierter Wirbel nicht durch die Krebserkrankung bedingt war (Torres & Hammond, 2016). Genauso schließt das Vorhandensein von bereits bestehenden osteoporotischen Frakturen in angrenzenden Wirbelkörpern die Entwicklung einer pathologischen Fraktur nicht aus (Torres & Hammond, 2016).

Wie gerade beschrieben können osteoporotische und pathologische Wirbelkörperfrakturen weder mithilfe ihres klinischen Erscheinungsbildes noch anamnestisch voneinander unterschieden werden. Zur sicheren Diagnostik ist man deshalb auf bildgebende Verfahren angewiesen: Die Diagnose einer Fraktur kann beispielsweise mithilfe von Röntgenaufnahmen, CT oder MRT gestellt werden. Die Differenzierung zwischen osteoporotischen und pathologischen Wirbelkörperfrakturen ist jedoch auf konventionellen Röntgenbildern aufgrund der unzureichenden technischen Qualität nicht möglich (Geith et al., 2015; Kendler et al., 2016; Torres & Hammond, 2016). Die Unterscheidung erfordert eine höherwertige Bildgebung (Torres & Hammond, 2016). Dazu gehört die CT und die MRT.

1.2.2.1 Computertomographie

Mit der Computertomografie (CT) lassen sich knöcherne Läsionen bei metastatischen Frakturen nachweisen, die Spezifität ist jedoch relativ gering (Geith et al., 2015).

Osteoporotische Wirbelkörperfakturen werden oftmals unterdiagnostiziert und bei computertomografischen Untersuchungen nicht erfasst (Tomita et al., 2018).

Im Folgenden sollen die CT-Merkmale, die bei benignen Wirbelkörperfrakturen mit statistischer Signifikanz häufiger gefunden werden, beschrieben werden: Frakturen der anterolateralen oder posterioren Kortikalis des Wirbelkörpers, ein retropulsierter Knochen und diffuse dünne paraspinale Weichteilverdickungen sprechen eher für eine benigne Wirbelkörperfraktur. Weiterhin ist das Puzzle-Zeichen ein Hinweis auf eine benigne Fraktur (Schwaiger et al., 2016). Unter dem Puzzle-Zeichen versteht man scharfe Bruchlinien ohne kortikale Zerstörung. Die verschobenen Knochenfragmente können in ihrer ursprünglichen Position rekonstruiert werden, um das Puzzle zu vervollständigen. Darüber hinaus kann das intravertebrale Vakuumphänomen, unter dem man einen luftgefüllten Spalt versteht, als Hinweis auf eine benigne Wirbelkörperfraktur angeführt werden. Dieses Phänomen wurde bei malignen Frakturen fast noch nicht beschrieben (Schwaiger et al., 2016).

Bei den CT-Befunden, die auf bösartige Wirbelkörperfrakturen schließen lassen, geht es vor allem um die Zerstörung von Strukturen und andererseits um Weichteilmassen. Jede Form der Zerstörung, sei es der Kortikalis, der Spongiosa oder des Pedikels, war prädiktiv für eine durch eine maligne Grunderkrankung ausgelöste Wirbelkörperfraktur (Schwaiger et al., 2016). Eine epidurale oder fokale paravertebrale Weichteilmasse, die normalerweise jedoch eher mithilfe einer MR-Bildgebung diagnostiziert wird, lässt auf eine bösartige Fraktur schließen (Mauch et al., 2018). Ein weiteres Zeichen einer pathologischen Wirbelkörperfraktur ist eine konvexe hintere Wirbelkontur (Schwaiger et al., 2016).

In der folgenden Tabelle 1 sind die wichtigsten Entscheidungskriterien zusammengefasst:

Modalität	Osteoporotische Wirbelkörperfrakturen	Pathologische Wirbelkörperfrakturen
CT	•Retropulsierter Knochen (Mauch et al., 2018; Schwaiger et al., 2016) •Puzzle-Zeichen (Mauch et al., 2018; Schwaiger et al., 2016) •scharfe Frakturlinien (Mauch et al., 2018) •intravertebrales Vakuum-Phänomen (Mauch et al., 2018; Schwaiger et al., 2016)	•Zerstörung von knöchernen Strukturen (Mauch et al., 2018; Schwaiger et al., 2016) •epidurale oder fokale paravertebrale Weichteilmasse (Mauch et al., 2018; Schwaiger et al., 2016) •konvexe hintere Wirbelkontur (Schwaiger et al., 2016)

Tabelle 1: Wichtigste Entscheidungskriterien zur Differenzierung von osteoporotischen und pathologischen Wirbelkörperfrakturen (CT)

1.2.2.2 Magnetresonanztomographie

Aufgrund des hohen Weichteilkontrasts ist die MRT derzeit die Bildgebung der Wahl, um gutartige von bösartigen Wirbelkörperfrakturen zu unterscheiden. Die MRT wird zur Beurteilung von Wirbelbrüchen in Betracht gezogen, wenn klinisch oder röntgenologisch der Verdacht auf eine bösartige Erkrankung oder Infektion besteht. Außerdem lässt sich das Rückenmark hier gut darstellen, was bei einer fraglichen Gefährdung des Rückenmarks hilfreich ist. Darüber hinaus kann ein in der MRT erkennbares Knochenödem auf eine frische Fraktur hinweisen. Dies ist hilfreich, wenn eine Vertebroplastie oder Kyphoplastie in Betracht gezogen wird und zwischen einer frischen oder alten Fraktur unterschieden werden muss (Kendler et al., 2016). Der genaue Zeitpunkt der Fraktur ist oftmals schwierig nur anhand der Anamnese der Patient*in oder der medizinischen Aufzeichnungen zu bestimmen (Mauch et al., 2018). Insbesondere

osteoporotische Frakturen sind mithilfe der MRT im Vergleich zu anderen Methoden der Bildgebung empfindlicher und spezifischer nachweisbar (Geith et al., 2015).

Im Normalfall kann gut zwischen pathologischen und osteoporotischen Frakturen unterschieden werden, wenn die Fraktur bereits über drei Monate alt ist.

Osteoporotische Wirbelkörperfrakturen weisen charakteristischerweise kleinere Bereiche mit meist linearer Signalveränderung mit geringer Signalintensität in T1- und T2-gewichteten Bildern auf (Jung et al., 2003). Zu den Kriterien für eine gutartige Fraktur gehören außerdem die Lokalisierung in der Brustwirbelsäule und eine unauffällige Bandanomalie, eine unauffällige Kontur und Fehlen einer Pedikelbeteiligung (Torres & Hammond, 2016). Bei Anwendung dieser Kriterien auf MRT-Scans konnte die Entität der Fraktur in der Studie mit 94%-iger Genauigkeit bestimmt werden (Moulopoulos et al., 1996). Weiterhin sind ein retropulsiertes hinteres Knochenfragment mit eckigen Rändern (Laredo et al., 1995), multiple Kompressionsfrakturen (Jung et al., 2003) und das Fehlen einer extraossären Masse (Torres & Hammond, 2016), sowie eine parallel zu den Endplatten verlaufende Frakturlinie (Torres & Hammond, 2016) als Zeichen für osteoporotische Wirbelkörperfrakturen zu werten.

Bei Osteoporose ist der zugrundeliegende Frakturmechanismus der Verlust der Knochenmineraldichte bei gleichzeitigem Erhalt des Knochenmarks. Die MRT kann in einzigartiger Weise das Knochenmark darstellen, das bei einer akuten osteoporotischen Fraktur vorübergehend ödematös wird. Eine normale, unauffällige Signalintensität des Knochenmarks des komprimierten Wirbelkörpers (Jung et al., 2003; Mauch et al., 2018; Moulopoulos et al., 1996) oder der zumindest teilweise Erhalt der normal hohen T1- und

intermediären T2-Signalintensität (Jung et al., 2003; Mauch et al., 2018; Torres & Hammond, 2016) des Knochenmarks sprechen für eine gutartige Wirbelkörperfraktur.

Zu den Kriterien einer pathologischen Wirbelkörperfraktur zählt eine abnorme MR-Signalintensität im Pedikel, was als spezifisch für eine bösartige Fraktur angesehen wird (Cuénod et al., 1996; Moulopoulos et al., 1996). Dies basiert auf der Beobachtung, dass sich Metastasen im Wirbelkörper entwickeln und sich sekundär in die Pedikel ausbreiten (Algra et al., 1992; Torres & Hammond, 2016). Nicht nur eine abnorme Signalintensität im Pedikel, sondern auch im Wirbelkörper ist ein Zeichen für eine pathologische Fraktur: Bei malignen Wirbelkörperfrakturen infiltriert der Tumor das Knochenmark und schließlich die Trabekel und die Kortikalis. Bei bösartigen metastasierten Wirbelkörperfrakturen ist die normalerweise hohe T1-Signalintensität des Knochenmarks oft vollständig verschwunden, was zu einer diffusen, homogenen Abnahme der Signalintensität in T1-gewichteten Bildern führt (Moulopoulos et al., 1996; Torres & Hammond, 2016). Mauch et al. beschreibt dieses Phänomen bei bis zu 68% der metastatischen Läsionen (Mauch et al., 2018). Die maligne Infiltration kann entweder fokal oder diffus sein. Es gibt keine spezifische Veränderung der Signalintensität für die verschiedenen zugrunde liegenden Pathologien. Sie zeigen sich grundsätzlich als hypointens auf T1-gewichteten Bildern und hyperintens auf T2-gewichteten und Short Tau Inversion Recovery (STIR)-Bildern (Karchevsky et al., 2008).

Ein weiteres Zeichen für eine pathologische Fraktur ist das Vorhandensein einer fokalen paraspinalen oder epiduralen Masse, dies weist auf eine extra-vertebrale Ausbreitung von Metastasen hin (Torres & Hammond, 2016). Ebenfalls kann von einer bösartigen

Fraktur ausgegangen werden, wenn eine konvexe hintere Kontur des Wirbelkörpers (Moulopoulos et al., 1996; Torres & Hammond, 2016) oder eine Lokalisation in der Lendenwirbelsäule vorliegt (Moulopoulos et al., 1996).
Sowohl gutartige als auch neoplastische Erkrankungen nehmen Kontrastmittel auf (Karchevsky et al., 2008). Schmorl-Knoten waren für keine der beiden Entitäten charakteristisch (Moulopoulos et al., 1996).

Die differenzialdiagnostische Abklärung von frischeren, jüngeren Frakturen ist oftmals nicht eindeutig möglich (Geith et al., 2015). Die Faktoren, die diese Differenzierung verkomplizieren, werden im folgenden Abschnitt erörtert.
Gutartige Wirbelkörperfrakturen weisen typischerweise ein bandförmiges Ödem mit geringer Signalintensität neben der kollabierten Endplatte in der T1-gewichteten Bildgebung auf, das die primäre Frakturlinie darstellt und normales Knochenmark im übrigen Wirbelkörper (Karchevsky et al., 2008). Manchmal ist das Ödem jedoch übermäßig groß und betrifft den gesamten Wirbelkörper. Dies wirft die Frage auf, ob im Wirbelkörper möglicherweise ein neoplastischer Prozess vorliegt (Karchevsky et al., 2008). Vor allem akute, weniger als zwei Wochen zurückliegende, und subakute, zwischen zwei und drei Wochen zurückliegende, gutartige Wirbelkörperfrakturen weisen oft große Bereiche mit MR-Signalveränderungen oder erhöhtem Stoffwechsel auf, der mithilfe von nuklearmedizinischen Modalitäten sichtbar gemacht werden können. Diese Veränderungen, zu denen intertrabekuläre Blutungen, Ödeme und frühe Reparationsprozesse gehören, können eine Malignität imitieren (Mauch et al., 2018).

Beispielsweise führen Ödemen zu einer diffusen Hypointensität in T1-gewichteten Sequenzen und zeigen eine fleckige Anreicherung (Mauch et al., 2018).

In der folgenden Tabelle 2 werden die wichtigsten Entscheidungskriterien zusammengefasst:

Modalität		Osteoporotische Wirbelkörperfrakturen	Pathologische Wirbelkörperfrakturen
MRT	Morphologie	•retropulsierte Knochenfragmente (Jung et al., 2003; Laredo et al., 1995) •keine Bandanomalie (Jung et al., 2003; Moulopoulos et al., 1996) •Normales posteriores Elementsignal (Moulopoulos et al., 1996) •zusätzliche gutartige VCFs (Jung et al., 2003)	•Pedikelbeteiligung (Cuénod et al., 1996; Jung et al., 2003; Moulopoulos et al., 1996; Torres & Hammond, 2016) •Konturanomalie (Jung et al., 2003; Moulopoulos et al., 1996) •Epidurale/ paravertebrale Weichteilmasse (Torres & Hammond, 2016) •konvexe hintere Wirbelkontur (Moulopoulos et al., 1996; Torres & Hammond, 2016) •Metastasen in anderen Wirbeln
	Signal und KM-Aufnahmemuster	•Erhaltenes normales Marksignal (Torres & Hammond, 2016) •Regelmäßige Ränder (Mauch et al., 2018) •Lineare Signalverstärkung (Geith et al., 2015) •Lineare horizontal hypointensiv T1/T2-Band (Jung et al., 2003; Torres & Hammond, 2016)	•Geographische Verdrängung des normalen Marksignals (Mauch et al., 2018) •unregelmäßige Ränder (Mauch et al., 2018) •Erhöhte Anreicherung im Vergleich zu benachbarten Wirbeln und nach 3 Monaten (Jung et al., 2003)
	Diffusion	•Keine eingeschränkte Diffusion (Mauch et al., 2018)	•Erhöhte eingeschränkte Diffusion (Mauch et al., 2018)

Tabelle 2: Wichtigste Entscheidungskriterien zur Differenzierung von osteoporotischen und pathologischen Wirbelkörperfrakturen (MRT)

Zusammenfassend ist zu sagen, dass in den meisten Fällen gutartige und bösartige Frakturen anhand charakteristischer morphologischer Kriterien in der MRT unterschieden werden können. Die Spezifität ist jedoch relativ gering bei akuten oder subakuten Frakturen mit ausgeprägtem Knochenmarksödem oder wenn eine Diagnostik mithilfe einer MRT nicht möglich ist und nur eine CT zur Verfügung steht (Geith et al., 2015).

1.2.3 Therapie und Management

Zu den allgemeinen schmerzlindernden Maßnahmen gehören kurzfristige Bettruhe und Schmerzlinderung mit Paracetamol, nichtsteroidalen Antirheumatika und Narkotika (Blasco et al., 2012; Kendler et al., 2016; Mudano et al., 2009). Physiotherapie kann Patient*innen angeboten werden, um die Schmerzen zu lindern und die Mobilität zu verbessern (Chow et al., 1989; Harrison et al., 1993; Kendler et al., 2016; Sinaki, 2012; Sinaki et al., 2002). Zusätzlich verringert Bewegung das spätere Frakturrisiko bei Osteoporose (Chow et al., 1989; Harrison et al., 1993; Kendler et al., 2016; Sinaki, 2012; Sinaki et al., 2002).

In der akuten Phase ist der Einsatz verschiedener Schmerztherapietechniken möglich. Dazu gehören unter anderem Ultraschallbehandlungen, Hydrotherapie, Kryotherapie, Wärmeanwendungen und Dehnungsübungen zur Verringerung von Muskelkrämpfen (Blasco et al., 2012; Kendler et al., 2016; Mudano et al., 2009).

Wirbelkörperfrakturen können außerdem mit Vertebroplastien und Kyphoplastien behandelt werden. Diese vertebralen Augmentationen sind umstritten. Sie stellen eine Behandlungsoption dar, wenn trotz einer mindestens sechswöchigen konservativen medizinischen Therapie Schmerzen bestehen oder neurologische Defizite vorhanden sind. Vertebroplastien und Kyphoplastien können die Schmerzen bei Wirbelbrüchen kurzfristig lindern (Cho et al., 2015), haben jedoch den Nachteil, dass es zu Komplikationen bei der Behandlung kommen kann und das Risiko eines Bruchs benachbarter Wirbel erhöht ist (Blasco et al., 2012; Kendler et al., 2016; Mudano et al., 2009). Grundsätzlich zielt die Behandlung darauf ab, die mechanische Instabilität und die Nervenkompression zu beheben, die primär zu den Symptomen führen. Aus diesem Grund wird von vielen Autor*innen eine chirurgische Dekompression und Stabilisierung mit Instrumentation als attraktive Behandlungsoption beschrieben. Viele Studien deuten auf gute klinische Ergebnisse nach einer chirurgischen Behandlung hin, einschließlich Schmerzkontrolle und Wiedererlangung oder Erhaltung der Mobilität (Hirabayashi et al., 2003; Ibrahim et al., 2008; Kwon et al., 2009; Yamashita et al., 2008).

Zusätzlich zu diesen allgemeinen Maßnahmen muss bei Patient*innen mit osteoporotischen Wirbelbrüchen eine frakturhemmende Therapie in Betracht gezogen werden. Denn sowohl symptomatische als auch asymptomatische Wirbelfrakturen weisen stark auf ein erhöhtes Frakturrisiko bei unbehandelten Patient*innen hin. In der Kohorte der *Study of Osteoporotic Fractures* hatten Frauen mit einer Wirbelfraktur in der medizinischen Vorgeschichte ein etwa dreifach höheres Risiko für eine neue Wirbelfraktur als Frauen ohne bereits vorbestehende Wirbelfraktur (Cauley et al., 2007).

Bei Patient*innen, die während einer klinischen Osteoporose-Studie ein Placebo erhielten und eine neue Wirbelfraktur erlitten, lag die Inzidenz einer weiteren neuen Wirbelfraktur innerhalb eines Jahres bei 20 % (Ferrar et al., 2012). Kurz nach einer Wirbelfraktur besteht ein signifikant erhöhtes Risiko für jede Art von Fraktur (Johnell et al., 2004; van Geel et al., 2010). Patient*innen sollten deshalb so bald wie möglich nach der Diagnose einer Fraktur eine geeignete Therapie erhalten. Zahlreiche pharmakologische Therapien verringern das Risiko einer Wirbelkörperfraktur erheblich (Black et al., 2007; Cummings et al., 2009; Kendler et al., 2016; MacLean et al., 2008). Eine pharmakologische Therapie der Osteoporose ist insbesondere Patient*innen mit jüngeren, höhergradigen oder mehrfachen Frakturen dringend zu empfehlen. Die größte absolute Risikoreduktion für künftige Frakturen kann bei Patient*innen mit einer Wirbelkörperfraktur in der Anamnese erzielt werden. Als hochwirksame pharmakologische Arzneimittel sind Teriparatid, Zoledronsäure oder Denosumab anzuführen. Sekundäre Ursachen für die Fraktur sollten vor Beginn der Therapie diagnostiziert und behandelt werden. Aufgrund des deutlich erhöhten zukünftigen Frakturrisikos nach einer Wirbelkörperfraktur betonen zahlreiche Praxisleitlinien die Bedeutung einer Pharmakotherapie zur Verringerung des Risikos zukünftiger Frakturen, unabhängig von der 10-Jahres-Frakturrisikobewertung (FRAX) oder der Knochenmineraldichte (Kendler et al., 2016; Papaioannou et al., 2010; The Board of Trustees of The North American Menopause Society, 2010).

Die Therapie von pathologischen Wirbelkörperfrakturen stellt sich ein wenig komplexer dar: Pathologische Frakturen sind ein bekanntes Problem für Krebspatient*innen, es gibt

jedoch derzeit keinen Konsens über die richtige Behandlungsstrategie (Cho et al., 2015). Als Behandlungsoptionen kommen neben den oben beschriebenen Therapieoptionen, zu denen eine Vertebroplastie (VP) oder eine Kyphoplastie (KP) und eine operative Versorgung (OP) gehören, auch eine Strahlentherapie (RT), eine Chemotherapie und weitere Behandlungsoptionen wie eine Steroidtherapie (Cho et al., 2015; Chong et al., 2012; Ha et al., 2015; Lim et al., 2009; Schuster & Grady, 2001). In einer koreanischen Studie mit 54 eingeschlossenen Patient*innen war die Strahlentherapie die häufigste primäre Behandlungsoption bei metastasierten pathologischen Frakturen der Wirbelsäule. Die Ansprechrate war jedoch suboptimal (Cho et al., 2015). In dieser Studie konnte außerdem bei Patient*innen, die mit einer palliativen Operation behandelt wurden, eine Verbesserung der Lebensqualität nachgewiesen werden (Cho et al., 2015).

Das Erstellen einer Leitlinie für die Diagnose und Behandlung von metastasierten pathologischen Frakturen der Wirbelsäule ist von großer Bedeutung.

1.3 Ziel der Arbeit

Die korrekte Differenzierung zwischen osteoporotischen und pathologischen Wirbelkörperfrakturen ist entscheidend für die weitere Vorgehensweise, Therapieoptionen und die Prognose der Patient*innen. Normalerweise können gutartige von bösartigen Frakturen durch charakteristische morphologische Kriterien in der MRT oder durch die Kombination von CT und MRT unterschieden werden. In Einzelfällen ist die Entität von insbesondere akuten und subakuten Frakturen, die mit einem ausgeprägten Knochenmarködem einhergehen, nicht klar zu bestimmen. Vor allem wenn keine MRT-Bildgebung vorhanden ist und nur CT-Bilddaten zur Verfügung stehen, ist die Differenzierung oftmals nicht eindeutig möglich.

Sowohl für Patient*innen, die sich nicht unnötigerweise weiteren Untersuchungen unterziehen müssten, als auch für das Gesundheitssystem und Radiolog*innen wäre es von Vorteil, basierend auf ausschließlich einer CT-Untersuchung sicher differenzieren zu können.

Aus diesem Grund ist es von großer Bedeutung, neue Algorithmen zu entwickeln und zu evaluieren, die Radiolog*innen bei der Diagnose unterstützen. Dadurch können individualisierte und angemessene Behandlungsentscheidungen getroffen werden.

Derzeit ist noch unklar, ob und wie genau Deep Learning Algorithmen zwischen osteoporotischen und pathologischen Frakturen unterscheiden können, wenn nur CT-Daten zur Verfügung stehen. Darüber hinaus stellt sich die Frage, ob die Leistung des Deep Learning Algorithmus der Leistung unerfahrener oder erfahrener Radiolog*innen gleichwertig oder sogar überlegen ist.

Ziel dieser Studie ist die Entwicklung und Validierung eines Deep Learning Algorithmus unter Verwendung von Methoden des maschinellen Lernens, der osteoporotische und pathologische Wirbelfrakturen auf CT-Bilddaten zuverlässig erkennen, bewerten und zwischen diesen unterscheiden kann. Dadurch soll die korrekte Diagnose von Wirbelfrakturen gewährleistet und eine patientenspezifische Therapieentscheidung getroffen werden.

2 Material und Methoden

2.1 Aufbereitung der Datenbank

Diese retrospektive, multizentrische Studie wurde von den lokalen institutionellen Prüfungsausschüssen genehmigt (Ethikausschuss Technische Universität München, Vorschlag Nr. 460/20S und Ethikausschuss Ludwig-Maximilians-Universität München, Vorschlag Nr. 21-0209). Die Studie wurde in Übereinstimmung mit nationalen und internationalen Richtlinien durchgeführt. Für diese retrospektive, anonymisierte Analyse wurde auf die schriftliche Einwilligung nach Aufklärung verzichtet.

Datensätze

Es wurden retrospektiv MDCT-In-vivo-Daten von Patient*innen mit mindestens einer osteoporotischen oder pathologischer Wirbelfraktur im institutionellen Bildarchivierungs- und Kommunikationssystem (PACS) des Universitätsklinikums der Technischen Universität München, Klinikum Rechts der Isar identifiziert. Es wurden MDCT-In-vivo-Daten eingeschlossen, die zwischen Juni 2005 und Dezember 2022 aufgenommen wurden. Die Untersuchung der Frakturen erfolgte auf vertebral spezifischer Ebene. Es wurden ausschließlich Patient*innen mit diagnostizierten Wirbelkörperfrakturen in der Brust- und Lendenwirbelsäule aufgenommen. Aufgrund der anatomischen Unterschiede der Wirbelkörper wurden zervikale und sakrale Wirbelkörperfrakturen ausgeschlossen. Es wurden nur Patient*innen mit mindestens einer präoperativen CT- und/oder MRT-Bildgebung berücksichtigt. Pathologische und osteoporotische Frakturen wurden mithilfe eines kombinierten Goldstandards definiert, der klinische und radiologische Informationen umfasst:

Eine Wirbelkörperfraktur wurde als bösartig definiert, wenn ein histopathologischer Nachweis einer neoplastischen Infiltration vorlag. Dieser wurde entweder durch eine Biopsie oder Operation gewonnen. Wenn kein histopathologischer Nachweis vorlag, musste die Fraktur von einer zertifizierten Radiolog*in mit Fachkenntnissen in der Wirbelsäulenbildgebung untersucht werden (A.S.G.). Die Fraktur wurde als pathologisch definiert, wenn typische bildgebende Befunde für neoplastische Informationen in zusätzlichen bildgebenden Verfahren, einschließlich SPECT-Bildgebung, MR-Bildgebung und PET/CT-Bildgebung zu finden waren. Darunter versteht man im Detail vertebrale Läsionen, die bei der Einzelphotonen-Emissions-Computertomographie (SPECT) verdächtig nach vertebralen Metastasen aussahen, Frakturen mit einer Nachuntersuchung, die eine Progredienz der neoplastischen Infiltration zeigten, MR-Bildgebungsbefunde, die von der Radiolog*in als typisch für eine bösartige Fraktur eingestuft wurden, oder erhöhte maximale Standardaufnahme-Werte (SUV), die bei einer vorherigen Positronen-Emissions-Tomographie (PET)/CT-Untersuchung des nicht gebrochenen Wirbels auf Wirbelmetastasen hindeuteten. SUV-Werte in der PET/CT-Bildgebung sind bei akuten und subakuten Frakturen typischerweise erhöht. Deshalb liefert die PET/CT-Bildgebung bei einem bereits gebrochenen Wirbel nur begrenzte Informationen und wurde daher nicht als Maß für die Definition einer bösartigen Fraktur herangezogen, wenn die Untersuchung nach dem Auftreten der Fraktur durchgeführt wurde (Shon & Fogelman, 2003).

Wirbelkörperfrakturen wurden ebenfalls als pathologisch eingestuft, wenn eine bildgebende Folgeuntersuchungen vorhanden war, die ein Fortschreiten der neoplastischen Infiltration zeigte.

Darüber hinaus wurden alle nicht gebrochenen Wirbel auf das Vorhandensein oder Nichtvorhandensein von bösartigen Läsionen untersucht. Wirbel mit unklaren Befunden wurden von der Analyse ausgeschlossen (n=25).

Um Wirbelkörper als osteoporotisch zu definieren, musste eine bildgebende Nachuntersuchung mindestens drei Monate nach der ursprünglichen Untersuchung vorliegen, um eine fortschreitende neoplastische Infiltration auszuschließen.

Für die abschließenden unabhängigen Tests und die geografische Validierung wurde eine externe Testgruppe von einem weiteren Universitätsklinikum, dem Klinikum Großhadern genutzt. Der externe Testsatz umfasste Patienten mit MDCT-Scans und Wirbelfrakturen. Hier wurde derselbe kombinierte Goldstandard verwendet, der für den internen Datensatz zur Definition osteoporotischer und bösartiger Frakturen und nicht frakturierter Wirbel mit bösartigen Läsionen angewandt wurde. Auch hier wurden Wirbel mit unklaren Befunden ausgeschlossen (n=1).

MDCT-Bildgebung

Die MDCT-Scans wurden mit 12 verschiedenen MDCT-Scannern durchgeführt (Brilliance 64, iCT 256, Ingenuity Core 128 und IQon, Philips Medical Systems; Siemens Biograph 64 und 128, Somatom Emotion 16, Sensation 16 und 64, Sensation Cardiac 64 und Somatom Definition AS und AS+, Siemens Healthineers). Alle Scans wurden im Helikalmodus mit einer Röhrenspitzenspannung von 120 kVp, einer axialen Schichtdicke von 0,9-1 mm und einer adaptiven Röhrenbelastung aufgenommen. Einige Aufnahmen wurden nach Verabreichung eines oralen (Barilux Scan, Sanochemia Diagnostics) und/oder intravenösen (Imeron 400, Bracco) Kontrastmittels durchgeführt. Die

Aufnahmen nach intravenöser Kontrastmittelgabe erfolgten in der arteriellen oder portalvenösen Phase. Der Start der Aufnahme wurde in diesen Fällen entweder ausgelöst durch einen Schwellenwert für die Kontrasterhöhung in der definierten Region, beispielsweise in der Aorta, oder nach einer bereits vorher festgelegten Verzögerung von 70 Sekunden.

Vorverarbeitung der Daten

In einem ersten Schritt erfolgte die Anonymisierung der MDCT-Daten, im Anschluss daran wurden die Daten in einem Dateiformat der Neuroimaging Informatics Technology Initiative (NifTi) extrahiert. Weiterhin wurden die Daten auf eine maximale isotope räumliche Auflösung von 1mm reduziert. In einem weiteren Schritt wurden die extrahierten Daten mit einer automatisierten, intern entwickelten Deep Learning Pipeline für die vollautomatische Wirbelsäulenbeschriftung und Wirbelsäulensegmentierung vorbereitet, wie zuvor beschrieben (Löffler et al., 2020; Sekuboyina et al., 2020, 2021). In dieser Pipeline werden drei künstliche neuronale Netze implementiert, um erstens die Wirbelsäule zu erkennen, zweitens jeden Wirbel zu kennzeichnen und drittens jeden Wirbel zu segmentieren (Abbildung 2).

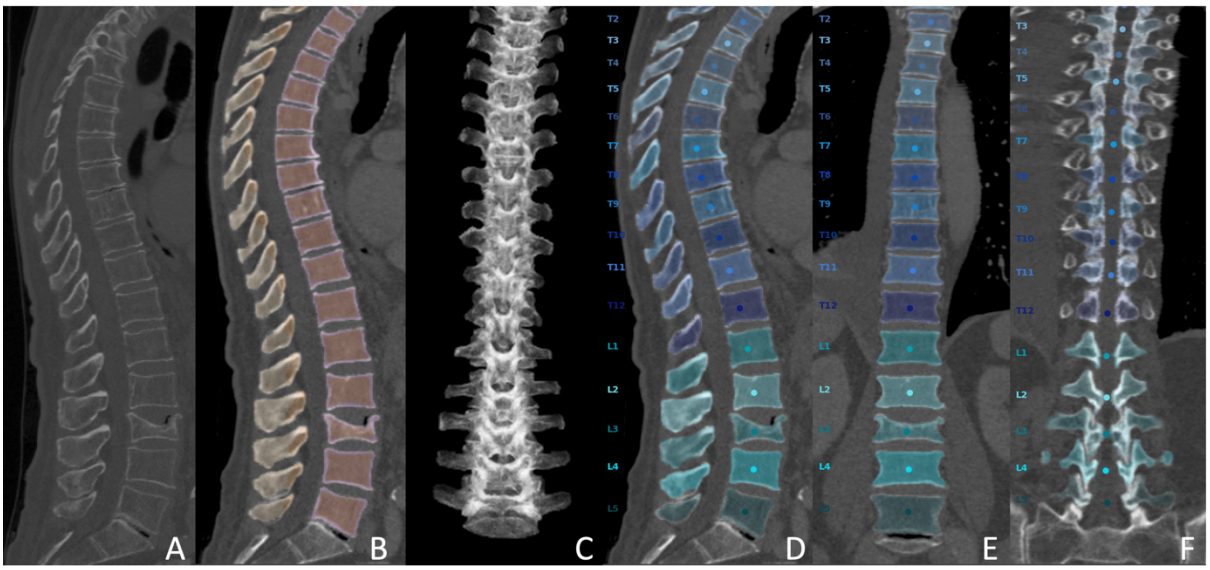

Abbildung 2: Vorverarbeitung der Daten. Überblick über die automatisierte, auf Deep Learning basierende Pipeline, die verwendet wurde. (Foreman et al., 2024)

A: Originaldaten; B und C: Segmentierung der Wirbelkörper; D-F: Beschriftung der Wirbelkörper.

Die vom automatisierten Deep Learning Modell beschrifteten und segmentierten Daten wurden anschließend überprüft. Bei Bedarf wurden manuelle Anpassungen vorgenommen. Auf der Grundlage der Segmentierungen der Wirbelsäule wurden Bounding Boxes erstellt. Darunter versteht man imaginäre Begrenzungsrahmen. Für jeden Wirbel wurden 3D-Felder der Größe 64x64x64 extrahiert.

2.2 Entwicklung und Training der Deep Learning Modelle

Vor dem Training wurden 18 % des internen Datensatzes nach dem Zufallsprinzip als interner Hold-out-Testsatz ausgewählt und vom restlichen Datensatz separiert. Die verbleibenden 82% wurden für das Training und die Validierung verwendet. Da die

Morphologie einer Fraktur bei ein und derselben Patient*in ähnlich sein kann, wurde die Datenaufteilung auf Proband*innenebene vorgenommen. Hiermit sollten Datenverluste vermieden werden. Es wurde nur eine Fraktur pro Patient*in im internen und externen Testsatz analysiert.

Die Deep Learning Modelle basierten auf einer 3D UNET Encoder-Classifier-Architektur. Alle Modelle wurden in PyTorch (1.7.0+cu101) mit einer 16-GB NVIDIA Quadro RTX 4000 (v537.42) entwickelt. Die Modelle wurden mit einer Stapelgröße von 32 und einer Lernrate von 1e-4 unter Verwendung eines Adam-Optimierers trainiert. Um das beste Modell auswählen zu können, wurde das Training wurde mit frühzeitigem Abbruch und Überwachung der Validierungsverluste durchgeführt. Als Verlustfunktion wurde die Cross-Entropie verwendet. Während des Trainings wurden Datenerweiterungen wie Links-Rechts-Spiegelung, zufällige sagittale Drehung (-10, +10 Grad), zufälliger Zoom und zufällige elastische Verformung des Gitters vorgenommen.

Für das Training und die Validierung der Modelle wurden zwei verschiedene Ansätze implementiert und ihre Leistung verglichen. Im ersten Ansatz wurden ausschließlich osteoporotische und maligne Wirbelfrakturen aus dem Trainings-/Validierungsdatensatz einbezogen. Im zweiten Ansatz wurden auch die nicht frakturierten Wirbel mit bösartigen Läsionen aus dem Trainings-/Validierungsdatensatz inkludiert. Diese Wirbelkörper bildeten nun eine Gruppe mit den Wirbelkörpern mit pathologischen Frakturen.

Modellbewertung

Aus jedem Kreuzvalidierungslauf wurde das Modell mit der besten Leistung für die abschließende Bewertung des internen und externen Testsatzes ausgewählt.

Um einen aussagekräftigen Vergleich ziehen zu können, wurden die MDCT-Bilder des internen Hold-out-Testsatzes und des externen Testsatzes von zwei Assistenzärzt*innen für Radiologie (M.C.M. und G.C.F. mit 2 bzw. 4 Jahren Erfahrung) und einer Fachärzt*in für Radiologie (M.R. mit 7 Jahren Erfahrung) ebenfalls ausgewertet. Für die Befundung hatten die Ärzt*innen ausschließlich die MDCT-Bilder zur Verfügung. Sie waren gegenüber allen anderen bildgebenden Verfahren (einschließlich MR, SPECT, PET/CT und Folgeuntersuchungen), sowie klinischen und histopathologischen Befunden verblindet.

2.3 Statistische Analyse

Die Berechnungen für die Modellmetriken wurden mit Scikit-learn (https://scikit-learn.org/stable/index.html, 1.3.2.) durchgeführt. Die Leistung der Modelle wurde anhand des internen Hold-Out-Testsatzes und des externen Testsatzes zur Unterscheidung von osteoporotischen und pathologischen Wirbelfrakturen beurteilt. Es wurde die Accuracy, Sensitivität, Spezifität und die Area under the curve (AUC, significance level $p<0.05$) der Receiver Operating Characteristic (ROC)-Kurve ermittelt und bewertet. Weiterhin wurde das 95%-Konfidenzintervall (CI) berechnet. Die inter- und intrareader Reproduzierbarkeit wurde bei 20 zufällig ausgewählten Patient*innen untersucht, die einzelnen Auswertungen lagen mindestens 30 Tage auseinander.

Es wurde eine dreistufige Ausgabe implementiert: "definitiv osteoporotisch", "definitiv bösartig" und "unsicher, bitte MRT durchführen", um das Modell im klinischen Alltag besser nutzen zu können. Diese Einteilung wurde mithilfe einer Schwellenwertbildung der angegebenen Wahrscheinlichkeit des Modells mit der besten Leistung abgeleitet. Die

Berechnung der Wahrscheinlichkeit jeder Klassenvorhersage wurde mittels stochastischem Gradientenabstieg als integraler Bestandteil des CNN verwendet. Ein Überblick über den Arbeitsablauf ist in Abbildung 3 dargestellt.

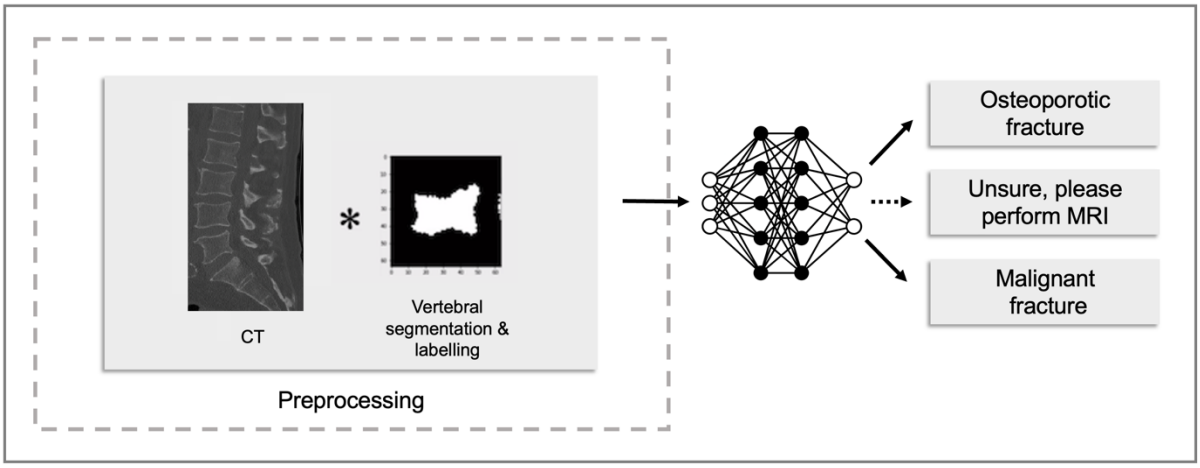

Abbildung 3: Überblick über den Arbeitsablauf. (Foreman et al., 2024)

Zunächst wurden die Bilddaten durch Segmentierung und Kennzeichnung der einzelnen Wirbel vorverarbeitet. In einem zweiten Schritt wurden die maskierten Bilder als Eingabe für die Deep Learning CNNs verwendet, und in Kategorien eingeteilt (gutartige oder bösartige Frakturen). Eine dritte Kategorie (unsicher, bitte MRT durchführen) wurde durch Schwellenwertbildung für die angegebene Wahrscheinlichkeit des besten Modells eingeführt.

3 Ergebnisse

3.1 Datenset

Insgesamt bestand das interne Datenset aus 467 Patient*innen. Das Durchschnittsalter betrug 69 Jahre (± 13,0 Jahre). Die Patient*innen waren zwischen 18 und 94 Jahre alt. 229 der Patient*innen waren weiblich (48% Frauen). Es wurden 415 als osteoporotische Frakturen, 496 als pathologische Frakturen und 482 als nicht gebrochene Wirbel mit bösartigen Läsionen unter Verwendung des kombinierten Goldstandards kategorisiert.

Nach Ausschluss von Wirbeln mit unklaren Befunden (n=25) wurden insgesamt 381 Patienten in den internen Trainings-Datensatz aufgenommen. Das Durchschnittsalter betrug 69,9 Jahre (±11,4 Jahre). 188 der Patient*innen waren weiblich (51% Frauen). Es wurden 378 als osteoporotische Frakturen, 447 als pathologische Frakturen und 482 Wirbel als nicht gebrochene Wirbel mit bösartigen Läsionen unter Verwendung des zusammengesetzten Referenzstandards kategorisiert.

Für die abschließende unabhängige interne Prüfung wurden 86 Patient*innen aus dem internen Datensatz in den internen Test-Datensatz aufgenommen. Das Durchschnittsalter betrug 66,9 Jahre (±12,0 Jahre). 41 Patientinnen waren weiblich (55% Frauen). In diesem Datensatz waren 37 Patient*innen mit osteoporotischen Frakturen und 49 Patient*innen mit pathologischen Frakturen.

In den externen Datensatz wurden insgesamt 65 Patient*innen aufgenommen. Das Durchschnittsalter betrug 68,8 Jahre (± 12,5 Jahre). Die Patient*innen waren zwischen 25 und 91 Jahre alt. 39 Patientinnen waren weiblich (60% Frauen). In diesem Datensatz waren 30 Patient*innen mit osteoporotischen Frakturen und 35 Patient*innen mit pathologischen Frakturen.

Die Merkmale der Proband*innen sind in Tabelle 3 zusammengefasst. Die Patient*innenmerkmale, Geschlecht und Alter, unterschieden sich nicht signifikant zwischen den Trainings-/Validierungs- und den beiden Testdatensätzen.

Datensatz	**Interner Datensatz n=467**	**Interner Trainings-/ Validierungssatz n=381**	**Interner Test-Datensatz n=86**	**Externer Test-Datensatz n=65**
Alter (Jahre)*	69,0 ± 13,0	69,9 ± 11,4	66,9 ± 12,0	68,8 ± 12,5
Altersspektrum (Jahre)	18-94	18-94	36-94	25-91
Geschlecht (Frauen)	229	188	41	39
Osteoporotische Frakturen	415	378	37	30
Pathologische Frakturen	496	447	49	35
Nicht-frakturierte Wirbelkörper mit malignen Läsionen	482	482	N/A	N/A
*Die Daten werden als Mittelwert ± Standardabweichung angegeben.				

Tabelle 3: Merkmale der Proband*innen

3.2 Leistung des Computertomographie Deep Learning Modells

Bewertung der entwickelten Modelle für maschinelles Lernen

Für den ersten Ansatz wurde das Deep Learning Modell ausschließlich mit osteoporotischen und pathologischen Frakturen trainiert. Dieses Modell erreichte eine AUC von 0,82 bei 78% Sensitivität, 86% Spezifität und 81% Accuracy auf dem internen Testsatz und eine AUC von 0,70 bei 74% Sensitivität, 67% Spezifität und 71% Accuracy auf dem externen Testsatz.

Für den zweiten Ansatz wurde ein separates Deep Learning Modell nicht nur mit osteoporotischen und pathologischen Frakturen trainiert, sondern zusätzlich wurden

nicht gebrochene Wirbel mit bösartigen Läsionen eingeschlossen. Dieses Modell erreichte im Vergleich zum ersten Ansatz eine höhere Sensitivität mit einer AUC von 0,85 bei 82% Sensitivität, 89% Spezifität und 85% Accuracy auf dem internen Testsatz und eine AUC von 0,75 bei 80% Sensitivität, 70% Spezifität und 75% Accuracy auf dem externen Testsatz.

Im dritten Ansatz wurde zusätzlich eine dreistufige Ausgabe ("definitiv osteoporotisch", "definitiv bösartig" und "unsicher, bitte MRT durchführen") implementiert, indem die bereitgestellte Wahrscheinlichkeit des besten Modells mit einem Schwellenwert versehen wurde. Dieses Modell erreichte eine höhere AUC von 0,91 bei 94% Sensitivität, 87% Spezifität und 91% Accuracy auf dem internen Testsatz und einer AUC von 0,76 bei 86% Sensitivität, 67% Spezifität und 78% Accuracy auf dem externen Testsatz.

Die Ergebnisse der Deep Learning Modelle sind in Tabelle 4 zusammengefasst.

Modell	**Score**	**Internes Test-Set**	**Externes Test-Set**
Erster Ansatz	**AUC**	**0,82 (CI: 0,74, 0,90)**	**0,70 (CI: 0,59, 0,81)**
	Accuracy	**81% (70/86; CI: 0,73, 0,90)**	**71% (46/65; CI: 0,60, 0,82)**
	Sensitivität	**78% (38/49; CI: 0,65, 0,89)**	**74% (26/35; CI: 0,59, 0,88)**
	Spezifität	**86% (32/37; CI: 0,74, 0,97)**	**67% (20/30; CI: 0,59, 0,83)**
Zweiter Ansatz	**AUC**	**0,85 (CI: 0,77, 0,92)**	**0,75 (CI: 0,64, 0,85)**
	Accuracy	**85% (73/86; CI: 0,77, 0,92)**	**75% (49/65; CI: 0,65, 0,85)**
	Sensitivität	**82% (40/49; CI: 0,70, 0,92)**	**80% (28/35; CI: 0,67, 0,92)**
	Spezifität	**89%**	**70%**

		(33/37; CI: 0,78, 0,98)	**(21/30; CI: 0,53, 0,86)**
Dritter Ansatz	**AUC**	**0,91 (CI: 0,83, 0,97)**	**0,76 (CI: 0,64, 0,88)**
	Accuracy	**91% (60/66; CI: 0,83, 0,97)**	**78% (39/50; CI: 0,66, 0,88)**
	Sensitivität	**94% (34/36; CI: 0,86, 1,00)**	**86% (25/29; CI: 0,72, 0,97)**
	Spezifität	**87% (26/30; CI: 0,73, 0,97)**	**67% (14/21; CI: 0,45, 0,86)**
Beim ersten Ansatz wurde das Deep Learning Modell nur mit osteoporotischen Frakturen bzw. bösartigen Frakturen trainiert. Beim zweiten Ansatz wurde das Modell mit allen osteoporotischen Frakturen im Vergleich zu bösartigen Frakturen in Kombination mit nicht frakturierten Wirbeln mit bösartigen Läsionen trainiert. Beim dritten Ansatz wurde eine 3-stufige Ausgabe ("definitiv osteoporotisch", "definitiv bösartig" und "unsicher, bitte MRT durchführen“) implementiert, indem die bereitgestellte Wahrscheinlichkeit des Modells mit der besten Leistung mit einem Schwellenwert versehen wurde. AUC = Area under the curve CI = 95% Konfidenzintervall			

Tabelle 4: Leistung der Deep Learning Modelle für das interne und externe Testset.

Die folgenden Abbildungen zeigen Beispiele für korrekte Klassifizierungen, für abweichende Einstufungen und Fehlklassifizierungen des Deep Learning Modells:

Abbildung 4 zeigt ein Beispiel für eine korrekte Klassifizierung einer pathologischen Fraktur mit der entsprechenden Vorhersagewahrscheinlichkeit des Deep Learning Modells.

Abbildung 5 zeigt ein Beispiel für eine korrekte Klassifizierung einer osteoporotischen Fraktur mit der entsprechenden Vorhersagewahrscheinlichkeit des Deep Learning Modells.

Abbildung 6 zeigt ein Beispiel für eine abweichende Kategorisierung des Deep Learning Modells im Vergleich zur Einstufung der Radiologen.

Abbildung 7 zeigt ein Beispiel für eine Fehlklassifizierung des Deep Learning Modells.

Abbildung 4: Beispiel für eine korrekte Klassifizierung einer pathologischen Fraktur mit der entsprechenden Vorhersagewahrscheinlichkeit des Deep Learning Modells. (Foreman et al., 2024)

A: Ein sagittales CT-Bild zeigt eine Fraktur eines Brustwirbelkörpers. Eine große osteolytische Läsion und ein konvexer hinterer Rand des frakturierten Wirbelkörpers sind

sichtbar. Das maschinelle Lernmodell stufte diese Fraktur mit einer Wahrscheinlichkeit von 100% als pathologisch ein. Die Assistenzärzte und der Facharzt für Radiologie stellten die gleiche Diagnose.

B und C: Die entsprechenden STIR- (B) und nativen T1-gewichteten (C) MR-Bilder zeigen ebenfalls typische Befunde für eine pathologische Fraktur. Hierzu gehören diffuse STIR-hyperintense und T1-hypointense Signalveränderungen des Knochenmarks im gesamten Wirbelkörper.

Die Diagnose wurde mithilfe einer anschließenden Biopsie histopathologisch bestätigt.

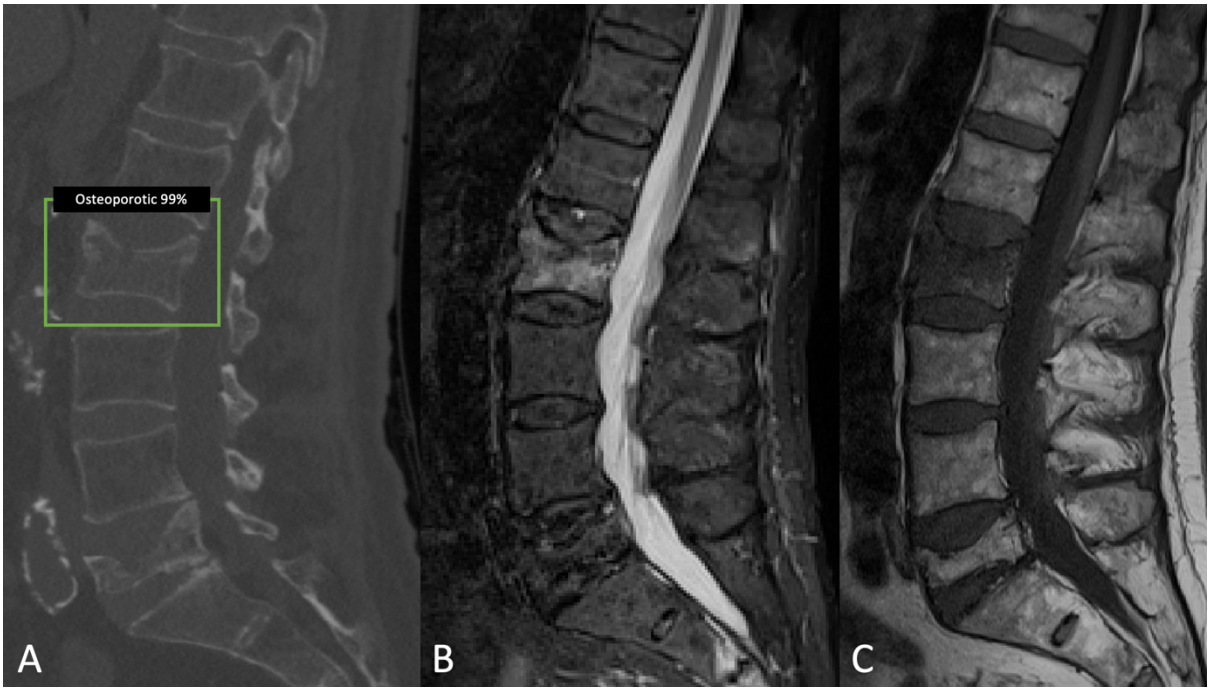

Abbildung 5: Beispiel für eine korrekte Klassifizierung einer osteoporotischen Fraktur mit der entsprechenden Vorhersagewahrscheinlichkeit des Deep Learning Modells. (Foreman et al., 2024)

A: Ein sagittales CT-Bild einer Fraktur der oberen Deckplatte. Das maschinelle Lernmodell stufte diese Fraktur mit einer Wahrscheinlichkeit von 99 % als osteoporotische Fraktur ein.

B und C: Die korrespondierenden MR-Bilder zeigen horizontale, bandförmige Signalveränderungen des Knochenmarks. Das normale Knochenmarkssignal bleibt in den nativen T1-gewichteteten Sequenzen (C) teilweise erhalten.

Die Diagnose wurde mithilfe von MR-Bildgebung und Nachuntersuchungsbefunden durch einen erfahrenen Facharzt für Radiologen bestätigt.

Abbildung 6: Beispiel für eine abweichende Kategorisierung des Deep Learning Modells im Vergleich zur Einstufung der Radiologen. (Foreman et al., 2024)

A: Ein sagittales CT-Bild einer Wirbelsäule. Es ist eine lumbale Wirbelfraktur erkennbar. Diese Fraktur wurde vom Deep Learning Modell mit einer Wahrscheinlichkeit von nur 56 % als osteoporotisch eingestuft, so dass diese Fraktur in die Kategorie "unsicher" fällt. Die Fraktur wurde von den Ärzten ebenfalls auf den CT-Scans als osteoporotisch eingestuft.

B: Die entsprechenden nativen T1-gewichteten Sequenzen zeigen diffuse hypointense Veränderungen des Knochenmarkssignals des Wirbelkörpers und multiple weitere hypointense Läsionen des Knochenmarks. Dies ist bildmorphologisch hochgradig verdächtig für eine bösartige Infiltration.

Die Diagnose einer pathologischen Fraktur wurde histopathologisch bestätigt.

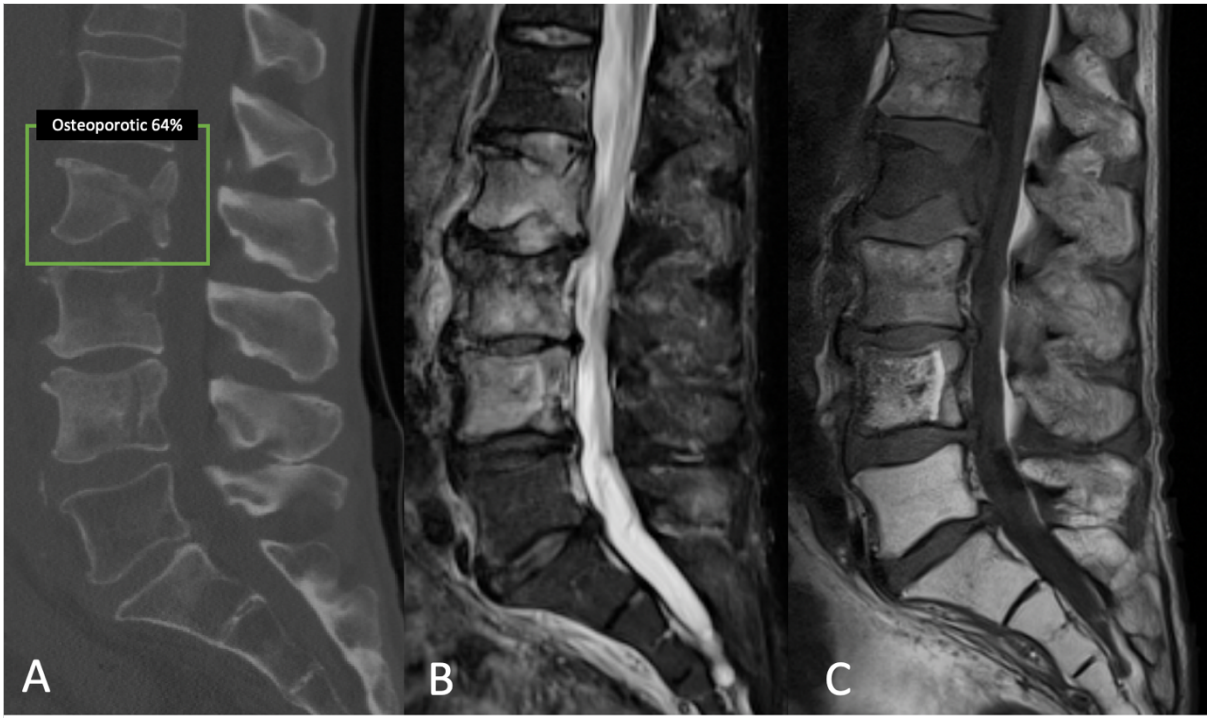

Abbildung 7: Beispiel für eine Fehlklassifizierung des Deep Learning Modells. (Foreman et al., 2024)

A: Ein sagittales CT-Bild eines Patienten mit bekanntem Prostatacarcinom und histologisch bestätigten ossären Metastasen. Das Deep Learning Modell klassifizierte die L2-Fraktur mit einer Wahrscheinlichkeit von 64 % als osteoporotisch.

B und C: Die entsprechenden STIR- (B) und nativen T1-gewichteten-(C) MR-Sequenzen zeigen homogene STIR-hyperintense, T1-hypointense

Knochenmarkssignalveränderungen des gesamten Wirbelkörpers. Dies ist hochgradig verdächtig für eine maligne Infiltration. Es sind weitere metastatische Läsionen in der Lendenwirbelsäule zu sehen. Weiterhin sind strahleninduzierte Signalveränderungen im fünften Lumbalwirbel und dem Kreuzbein nach einer vorangegangenen Strahlentherapie sichtbar.

3.4 Vergleich des Deep Learning Modells mit der Leistung von Radiolog*innen

Die Assistenzärzt*in mit 2 Jahren Erfahrung erreichte eine AUC von 0,69 bei 67% Sensitivität, 70% Spezifität und 69% Accuracy im internen Test-Set und im externen Test-Set eine AUC von 0,70 bei 63% Sensitivität, 77% Spezifität und 69% Accuracy.

Die Assistenzärzt*in mit 4 Jahren Erfahrung erreichte im internen Test-Set eine AUC von 0,71 bei 69% Sensitivität, 73% Spezifität und 71% Accuracy und im externen Test-Set eine AUC von 0,71 bei 63% Sensitivität, 80% Spezifität und 71% Accuracy.

Die zertifizierte Fachärzt*in für Radiologie erreichte eine AUC von 0,86 bei 94% Sensitivität, 78% Spezifität und 87% Accuracy im internen Testsatz und eine AUC von 0,71 bei 86% Sensitivität, 57% Spezifität und 72% Accuracy im externen Testsatz.

Das leistungsstärkste Deep Learning Modell erreichte im internen Test-Set eine AUC von 0,91 bei 94% Sensitivität, 87% Spezifität und 91% Accuracy und im externen Test-Set eine AUC von 0,76 bei 86% Sensitivität, 67% Spezifität und 78% Accuracy.

Die jeweiligen Ergebnisse sind in Tabelle 5 dargestellt.

Test-Set	**Score**	**Assistenz*ärztin mit 2 Jahren BE**	**Assistenz*ärztin mit 4 Jahren BE**	**Zertifizierte Radiolog*in mit 7 Jahren BE**	**Leistungsstärkstes Modell**
Interes	**AUC**	**0,69 (CI: 0,59, 0,78)****	**0,71 (CI: 0,61, 0,80)****	**0,86 (CI: 0,78, 0,93)**	**0,91 (CI: 0,83, 0,97)**

	Accuracy	**69% (59/86; CI: 0,59, 0,78)**	**71% (61/86; CI: 0,60, 0,80)**	**87% (75/86; CI: 0,80, 0,94)**	**91% (60/66; CI: 0,83, 0,97)**
	Sensitivität	**67% (33/49; CI: 0,54, 0,80)**	**69% (34/49; CI: 0,56, 0,82)**	**94% (46/49; CI: 0,86, 1,00)**	**94% (34/36; CI: 0,86, 1,00)**
	Spezifität	**70% (26/37; CI: 0,55, 0,85)**	**73% (27/37; CI: 0,58, 0,87)**	**78% (29/37; CI: 0,64, 0,91)**	**87% (26/30; CI: 0,73, 0,97)**
Externes	**AUC**	**0,70 (CI: 0,58, 0,80)**	**0,71 (CI: 0,60, 0,82)**	**0,71 (CI: 0,60, 0,82)**	**0,76 (CI: 0,64, 0,88)**
	Accuracy	**69% (45/65; CI: 0,58, 0,80)**	**71% (46/65; CI: 0,60, 0,82)**	**72% (47/65; CI: 0,62, 0,83)**	**78% (39/50; CI: 0,66, 0,88)**
	Sensitivität	**63% (22/35; CI: 0,46, 0,79)**	**63% (22/35; CI: 0,47, 0,78)**	**86% (30/35; CI: 0,73, 0,97)**	**86% (25/29; CI: 0,72, 0,97)**
	Spezifität	**77% (23/30; CI: 0,61, 0,91)**	**80% (24/30; CI: 0,65, 0,93)**	**57% (17/30; CI: 0,38, 0,74)**	**67% (14/21; CI: 0,45, 0,86)**

AUC = Area under the Curve, CI = 95% Konfidenzintervall,
*p ≤0,05 im Vergleich zum leistungsstärksten Deep Learning Modell,
**p ≤0,001 im Vergleich zum leistungsstärksten Deep Learning Modell
BE = Berufserfahrung

Tabelle 5: Leistung der Assistenzärzt*in für Radiologie mit 2 bzw. 4 Jahren Erfahrung und einer zertifizierten Radiolog*in mit 7 Jahren Erfahrung und Leistung des leistungsstärksten Modells. Die Leser waren gegenüber allen klinischen und histopathologischen Befunden verblindet.

4 Diskussion

Die Prävalenz von Wirbelkörperfrakturen ist in der heutigen immer älter werdenden Gesellschaft ansteigend. Um eine angemessene und patient*innenspezifische Therapie von Wirbelkörperfrakturen gewährleisten zu können, muss zwischen osteoporotischen und pathologischen Wirbelkörperfrakturen unterschieden werden. Wenn ausschließlich CT-Scans zur Verfügung stehen, ist die Differenzierung oftmals nur schwer möglich. Derzeit muss normalerweise eine kosten- und zeitaufwendige MRT-Bildgebung im Anschluss nach einer CT-Bildgebung durchgeführt werden, um zwischen den Differentialdiagnosen sicher entscheiden zu können. In dieser Arbeit wurden Deep Learning Algorithmen zur Unterscheidung von osteoporotischen und pathologischen Frakturen auf MDCT-Bildern entwickelt und validiert. In einem zweiten Schritt wurde die Leistung des Deep Learning Modells mit der Leistung von Radiolog*innen mit unterschiedlicher Berufserfahrung verglichen. Um die externe Validität zu garantieren, wurde zusätzlich ein externes Datenset evaluiert.

Die Ergebnisse des Modells mit der besten Leistung waren ausgesprochen gut.

Für den ersten Ansatz wurde das Deep Learning Modell ausschließlich mit osteoporotischen und pathologischen Frakturen trainiert. Dieses Modell erreichte eine AUC von 0,82 bei 78% Sensitivität, 86% Spezifität und einer Accuracy von 81% auf dem internen Testsatz und eine AUC von 0,70 bei 74% Sensitivität, 67% Spezifität und einer Accuracy von 71% auf dem externen Testsatz.

Interessanterweise verbesserte sich die Leistung unseres Modells bei der Unterscheidung zwischen osteoporotischen und pathologischen Wirbelfrakturen

erheblich, wenn zusätzlich zu den Wirbeln mit bösartigen Frakturen auch Wirbel ohne Frakturen mit malignen Läsionen in den Trainingsprozess mit einbezogen wurden. Durch die Inkludierung von Wirbelkörpern, die nicht frakturiert, aber in denen pathologische Läsionen sichtbar waren, konnte die Leistung des Modells verbessert werden. Dieses Modell erreichte im Vergleich zum ersten Ansatz eine höhere Sensitivität mit einer AUC von 0,85 bei 82% Sensitivität, 89% Spezifität und einer Accuracy von 85% auf dem internen Testsatz und eine AUC von 0,75 bei 80% Sensitivität, 70% Spezifität und einer Accuracy von 75% auf dem externen Testsatz. Diese Ergebnisse stehen im Einklang mit früheren Studien, in denen die Leistung von Deep Learning Modellen durch die Gruppierung von Bildern mit ähnlichem Inhalt verbessert werden konnte.(Husseini et al., 2020) Es wird davon ausgegangen, dass die Überschneidung charakteristischer Bildgebungsbefunde von nicht gebrochenen Wirbeln mit bösartigen Läsionen und bösartigen Frakturen den Trainingsprozess bereichert hat. Daraus folgte eine exaktere binäre Klassenvorhersage. Weiterhin ist die deutliche Zunahme der verfügbaren Trainingsdaten ein wichtiger Faktor, der zu besseren und zuverlässigeren Ergebnissen führt.

Durch die Implementierung einer dreistufigen Ausgabe, d. h. eine Einteilung in "definitiv osteoporotisch", "definitiv bösartig" und "unsicher, bitte MRT durchführen", konnte die Leistung des Modells weiter gesteigert werden. Dieses Modell erreichte eine höhere AUC von 0,91 bei 94% Sensitivität, 87% Spezifität und einer Accuracy von 91% auf dem internen Testsatz und einer AUC von 0,76 bei 86% Sensitivität, 67% Spezifität und einer Accuracy von 78% auf dem externen Testsatz.

Es ist bemerkenswert, wie gut der Deep Learning Algorithmus im Vergleich mit Assistenz- und Fachärzt*innen für Radiologie abschnitt: Die Leistung des Deep Learning Algorithmus war besser als die von Assistenzärzt*innen für Radiologie und vergleichbar mit der Leistung einer zertifizierten Radiolog*in, obwohl die Unterschiede nur für den internen Testsatz signifikant waren ($p<0,001$).

Diese Ergebnisse entsprechen den Ergebnissen aus früheren Studien. In den letzten Jahren wurden bereits Algorithmen entwickelt, die sich mit der Differenzierung zwischen osteoporotischen und pathologischen Wirbelkörperfrakturen befassten:
Es wurde bereits im Jahr 2021 von Li et al. ein Modell für maschinelles Lernen entwickelt und vorgestellt, um gutartige von bösartigen Wirbelfrakturen zu unterscheiden. Der Datensatz, mit dem dieser Algorithmus trainiert wurde, war jedoch wesentlich kleiner. 433 Patient*innen mit 296 bestätigten bösartigen und 137 gutartigen Frakturen wurden in diese retrospektive Studie eingeschlossen. Anders als in der hier vorgestellten Studie wurde das Modell anhand von drei manuell ausgewählten aufeinanderfolgenden CT-Schichten trainiert und validiert. Es wäre von Vorteil, die Schichtauswahl zu automatisieren (Li et al., 2021). Auch in der Studie von Li et al. wurden die Ergebnisse des Deep Learning Modells mit den Ergebnissen von Radiolog*innen mit unterschiedlich langer Berufserfahrung verglichen. Die drei Radiolog*innen mit jeweils 5, 3 und 1 Jahr Berufserfahrung erreichten eine Accuracy von 99 %, 95,2 % bzw. 92,8 %. In der ResNet50-Analyse lagen die diagnostische Sensitivität, Spezifität und Accuracy pro Schicht bei 90%, 79% und 85%. Wurden die Slices zu einer Diagnose pro Patient*in kombiniert, betrugen die Sensitivität, Spezifität und Accuracy 95%, 80% und 88% (Li et

al., 2021). Dieses Deep Learning Modell konnte also nicht auf gleicher Ebene wie eine erfahrene Radiolog*in zwischen osteoporotischen und pathologischen Frakturen differenzieren. Ein möglicher Grund hierfür ist der im Vergleich zu unserer Studie kleinere Datensatz, mit dem das Modell trainiert wurde. Eine große Einschränkung dieser Studie besteht jedoch darin, dass kein unabhängiger interner oder externer Testsatz zum Vergleich analysiert wurde. Es wurde nicht überprüft, ob die Ergebnisse mit anderen Datensätzen reproduzierbar wären (Li et al., 2021).

In einer weiteren Studie von Park et al. wurde ein CT-basiertes Radiomics Modell zur Unterscheidung von gutartigen und bösartigen Wirbelfrakturen beschrieben. Das Modell wurde mit 341 gebrochenen Wirbeln von insgesamt 158 Patient*innen trainiert. Die Ergebnisse wurden mit einem unabhängigen internen und externen Testsatz verglichen (Park et al., 2022). Der Algorithmus konnte auf ähnlichem Niveau wie eine Radiolog*in zwischen osteoporotischen und pathologischen Frakturen unterscheiden. Im externen Test-Set wurde für das Modell eine AUC von 0.83 und für die Radiolog*in eine AUC von 0.8 beschrieben. Der Unterschied zwischen der Leistung des Modells und der Leistung der Radiolog*innen war nicht statistisch signifikant. (p=0.37) Dieses Modell lieferte im Vergleich zu unserer Studie schlechtere Ergebnisse mit einer durchschnittlichen AUC von 0,84 auf dem internen Testsatz und 0,83 auf dem externen Testsatz. Diese Ergebnisse bestätigten die bisher publizierten Ergebnisse aus früheren Studien: Schon früher wurde gezeigt, dass CNN-Modelle normalerweise eine bessere Gesamtleistung im Vergleich zu Radiomics-Modellen aufweisen (Sun et al., 2020; Truhn et al., 2019). Ein weiterer entscheidender Grund für die schlechtere Ergebnisse könnte der im Vergleich zu der hier vorgestellten Studie kleinere Datensatz sein.

Es gibt weitere Studien, die auf anderen Bildgebungsmodalitäten basieren: Beispielsweise wurde von Filograna et al. ein Modell vorgestellt, das erfolgreich metastatische von nicht-metastatischen Wirbelkörpern in MRT-Scans unterscheiden konnte. Es wurden sowohl T1-gewichtete als auch T2-gewichtete Sequenzen genutzt. Die wichtigsten Prädiktoren basierten auf der T2-gewichteten Sequenz und waren morphologische und texturelle Merkmale (Filograna et al., 2019).

Zusätzlich konnte ein weiteres maschinelles Lernmodell erfolgreich zwischen gutartigen und bösartigen Knochenläsionen auf konventionellen Röntgenaufnahmen unterscheiden. Um die bestmögliche Leistung zu erreichen, wurden in dieses Modell sowohl radiologische Merkmale mithilfe der Röntgenaufnahmen, als auch demografische Informationen eingespeist. Dieses Modell war auf muskuloskelettale Tumore spezialisierten Radiolog*innen bezüglich der Accuracy unterlegen, aber Assistenzärzt*innen überlegen. Dies unterstreicht die Bedeutung der Entwicklung umfassender Modelle des maschinellen Lernens (von Schacky et al., 2022).

Einschränkungen

Es sind mehrere Einschränkungen bei der Interpretation dieser Studie zu beachten. Erstens wurde ein retrospektives Studiendesign verwendet, das grundsätzlich potenziell immer mit Verzerrungen bei der Auswahl des Patient*innenkollektivs verbunden ist. Weiterhin ist anzuführen, dass jegliche klinischen Informationen, von denen bekannt ist, dass sie sich auf die Knochengesundheit auswirken, wie Gewicht und Größe, Medikation oder Raucherstatus, nicht berücksichtigt wurden. Eine weitere Einschränkung ist, dass die entwickelten Modelle ausschließlich zwischen osteoporotischen und pathologischen

Wirbelfrakturen unterschieden und weitere Differentialdiagnosen wie beispielsweise entzündliche Erkrankungen nicht berücksichtigt wurden.

Zu den positiven Aspekten der aktuellen Studie gehört ihr multizentrisches Design, das die Bewertung der Modelle anhand eines unabhängigen externen Testsatzes ermöglicht hat. Dadurch konnte die externe Validität bewiesen werden. Außerdem ist dieser Datensatz derzeit der größte CT-Datensatz, der zum Training eines Deep Learning Modells zur Differenzierung von osteoporotischen und pathologischen Frakturen verwendet wurde. Ein weiterer positiver Aspekt ist die automatische Beschriftung der Wirbelkörper, die vor dem Training vorgenommen wurde. Es wurde bereits gezeigt, dass dies die Leistung von Deep Learning Modellen aufgrund der großen Formunterschiede zwischen den verschiedenen Wirbeln (Hals- und Lendenwirbel) nachweislich verbessert (Husseini et al., 2020). Weiterhin wurde der interne Datensatz für das Training und den Test auf Patient*innenebene aufgeteilt und nur eine Fraktur pro Patient*in in den Testsatz aufgenommen. Dadurch sollte vermieden werden, dass der Algorithmus an mehreren Frakturen derselben Patient*innen mit potenziell ähnlichen Bildgebungsbefunden getestet werden würde. Somit wäre die Stichprobe potenziell nicht repräsentativ und die Ergebnisse verzerrt.

Integration in den klinischen Alltag und zukünftige Forschung

Die Implementierung des Deep Learning CNNs zur Differenzierung zwischen osteoporotischen und pathologischen Wirbelkörperfrakturen in den klinischen Alltag würde zu einer Ersparnis von Zeit und monetären Ressourcen führen. Durch eine sichere Befundung von Wirbelkörperfrakturen und eine klare Einteilung in „definitiv

osteoporotisch", „definitiv bösartig" und „unsicher, bitte MRT durchführen", könnten Patient*innen nicht notwendige Untersuchungen und Verunsicherung bis zur endgültigen Diagnose erspart bleiben. Durch die Reduzierung der Anzahl von nicht notwendigen MRTs würde das Gesundheitssystem entlastet werden. Ressourcen könnten anderweitig und sinnvoller genutzt werden. Gleichzeitig würde den Radiolog*innen ein Teil der Arbeitsbelastung abgenommen werden.

Eine wichtige Frage ist wie ein Algorithmus erfolgreich in den klinischen Alltag integriert werden kann. Der hier vorgestellte Algorithmus könnte gut in einem klinischen Umfeld eingesetzt werden. Der Ärzt*in könnte die Differenzierung zwischen osteoporotischer und pathologischer Fraktur erleichtert werden, um das weitere Vorgehen zu bestimmen. Der Algorithmus könnte helfen die Entscheidung zu treffen, ob eine Patient*in aus der Notaufnahme mit Schmerzmitteln entlassen werden kann oder ob weitere diagnostische und klinische Untersuchungen erforderlich sind. Vor allem wenn die diensthabende Ärzt*in nur über begrenzte Erfahrung in der CT-Bildgebung verfügt, könnte dieser Algorithmus eine wertvolle Unterstützung sein.

Bezüglich zukünftiger Forschungsprojekte zur Differenzierung von osteoporotischen und pathologischen Wirbelkörperfrakturen ist Folgendes zu sagen: Weiterführende Studien sollten der Frage nachgehen, ob Modelle, die mit MRT-Daten trainiert wurden, bessere Ergebnisse liefern als Modelle, mit die CT-Daten trainiert wurden. Weiterhin wäre es interessant zu untersuchen, wie ein Modell abschneidet, das sowohl auf CT-, als auch auf MRT-Daten beruht. Dies setzt voraus, dass Patienten sich weiterhin zusätzlich einer MRT-Bildgebung unterziehen. Man erspart in diesem Fall den Patienten keine Untersuchung und dem Gesundheitssystem keine Kosten, trotzdem könnte es von Vorteil

sein, falls das Modell exzellente Leistungen erbringt. Wenn die Entität der Fraktur mit fast 100%-iger Wahrscheinlichkeit klassifiziert werden kann, könnte den Patienten ein invasiver Eingriff zur Gewebeentnahme erspart bleiben. Weiterhin könnte so eine flächendeckende hochqualitative Versorgung gewährleistet werden. Dies ist ein wichtiger Punkt im Zuge des zukünftigen Ärztemangels, der in den nächsten Jahren erwartet wird. Eine weitere ungeklärte Frage ist, wie erfolgreich ein Modell wäre, das Radiomics und CNN-Algorithmen kombiniert. Radiomics ist eine Methode, die eine große Menge an vordefinierten quantitativen Merkmalen aus Bildern extrahiert, die über Details hinausgehen, die für das menschliche Auge nicht erfassbar sind. Deep Learning beurteilt das gesamte Bild, Radiomics Modelle beurteilen nur einen Ausschnitt. Deep Learning CNNs und Radiomics Modelle könnten also theoretisch komplementäre bildgebende Biomarker liefern (Kalmet et al., 2020). Es konnte bereits in einer Studie, in der über Glioblastome geforscht wurde, gezeigt werden, dass durch die Kombinierung bessere Ergebnisse erzielt werden können, als durch alleinige Radiomics oder CNN Modelle (Calabrese et al., 2022).

Weiterhin wäre interessant zu eruieren, ob sich die Genauigkeit der Befunde von Radiolog*innen verbessert, wenn sie Zugriff auf eine Verdachtsdiagnose eines Deep Learning Modells haben. Dies konnte schon im Rahmen einer Studie bewiesen werden, in der Frakturen der Hüfte erfolgreicher diagnostiziert wurden, wenn die Assistenzärzt*innen für Radiologie von einem CNN-Algorithmus unterstützt wurden (Sato et al., 2021). Klinisch sehr interessant wäre ein CNN, das als Hilfestellung im klinischen Alltag beim Befunden der Wirbelkörperfrakturen für junge Radiolog*innen agiert. Dadurch könnte weiterhin ein Lerneffekt festgestellt werden, wenn die Radiolog*innen in

Ausbildung ihre initiale Bewertung der Fraktur mit der Bewertung des CNNs vergleichen könnten.

Klinische Faktoren liefern wichtige Informationen. Für die Zukunft wird es außerdem noch entscheidend sein, diese Informationen mit in die Entscheidungsfindung einzubeziehen. Nur so kann man ein vollständiges Bild erhalten und mit größerer Wahrscheinlichkeit die richtige Diagnose gestellt werden.

Zusammenfassend ist zu sagen, dass weitere Studien mit prospektiv erhobenen Daten gerechtfertigt sind. Es sollten mehr klinische Informationen berücksichtigt werden und mehr Differentialdiagnosen, beispielsweise entzündliche Erkrankungen wie Spondylodiszitis miteingeschlossen werden.

5 Zusammenfassung

5.1. Zusammenfassung auf Deutsch

Wirbelkörperfrakturen sind eine wachsende Herausforderung in unserer alternden Gesellschaft. Es wird sowohl ein Anstieg an osteoporotischen als auch ein Anstieg an pathologischen Wirbelkörperfrakturen verzeichnet. Die Unterscheidung zwischen osteoporotischen und pathologischen Wirbelkörperfrakturen ist von großer klinischer Bedeutung. Da sich die Behandlungsmöglichkeiten für die beiden Entitäten grundlegend unterscheiden – die Behandlung der zugrunde liegenden Osteoporose einerseits und die Behandlung der zugrunde liegenden bösartigen Erkrankung andererseits - ist eine schnelle und genaue Diagnose von entscheidender Bedeutung, um jeder Patient*in die erforderliche Behandlung zukommen lassen zu können.

Derzeit wird die Unterscheidung zwischen osteoporotischen und pathologischen Frakturen in der Regel mithilfe einer MRT durchgeführt. Mit der CT allein kann oftmals keine zuverlässige Diagnose gestellt werden. Es ist daher von großem Interesse, eine Methode zu entwickeln, die eine Diagnose anhand von kostengünstigen und schnellen CT-Bilddaten ermöglicht. Dies käme den Patient*innen und dem Gesundheitssystem zugute.

Insbesondere Deep Learning Modelle haben in den letzten Jahren hervorragende Ergebnisse erzielt, vor allem im Bereich der medizinischen Bildgebung. In dieser Studie wurde ein Deep Learning Modell erfolgreich trainiert und validiert, das eine hohe Trennschärfe bei der Unterscheidung zwischen osteoporotischen und pathologischen Frakturen auf MDCT-Bildern aufweist. Nachdem in das Modell einerseits osteoporotische und andererseits pathologische Wirbelkörperfrakturen und nicht frakturierte Wirbelkörper

mit malignen Läsionen inkludiert wurden und eine dreistufige Ausgabe implementiert wurde, konnten sehr gute Ergebnisse erzielt werden. Die Performance war dabei besser im Vergleich zu Assistenzärzt*innen der Radiologie und vergleichbar mit Experten-Radiolog*innen. Deep Learning Modelle, wie das hier vorgestellte, haben das Potenzial, die Arbeitsbelastung von Radiolog*innen zu verringern, den Einsatz von Ressourcen zu optimieren und eine angemessene und rechtzeitige Versorgung der Patient*innen zu gewährleisten. Langfristig soll eine KI-gestützte, qualitativ hochwertige Bildanalyse ermöglicht werden.

Mit Hilfe des hier vorgestellten Deep Learning Modells könnte die Zahl der für die endgültige Diagnose erforderlichen zusätzlichen Untersuchungen reduziert werden.

Weitere Studien sind erforderlich, um Algorithmen zu entwickeln, die erfolgreich in die klinische Routine integriert werden können.

5.2 Zusammenfassung auf Englisch

Vertebral fractures are a growing challenge of our ageing society. There is an increase in both osteoporotic and pathological vertebral body fractures. The differentiation between osteoporotic and pathological vertebral body fractures is of great clinical importance. Since the treatment options for the two entities are fundamentally different - treatment of the underlying osteoporosis on the one hand and treatment of the underlying malignant disease on the other - a rapid and accurate diagnosis is crucial in order to be able to provide each patient with the necessary treatment.

Currently, the differentiation of osteoporotic and pathological fractures is usually carried out using an MRI. It is often not possible to make a reliable diagnosis using CT alone. It is therefore of great interest to develop a method of making a diagnosis using inexpensive and fast CT image data. This would benefit patients and the healthcare system.

Deep learning models in particular have produced excellent results in recent years, especially in the field of medical imaging. In this study, a deep learning model was successfully trained and validated, which has a high discriminatory power in differentiating between osteoporotic and pathological fractures on MDCT images. After including osteoporotic and pathological vertebral fractures on the one hand and non-fractured vertebral bodies with malignant lesions on the other and implementing a three-stage output, very good results were achieved. The performance was higher than the performance of radiology residents and comparable to expert radiologists.

Deep learning models, such as the one presented here, have the potential to reduce the workload of radiologists, optimize the use of resources and ensure appropriate and timely

care for patients. In the long term, the aim is to provide AI-enhanced high-quality imaging analysis.

With the help of the deep learning model presented here, the number of additional examinations required for the final diagnosis could be reduced.

Further studies are needed to develop algorithms that can be successfully integrated into the clinical routine.

6 Literaturverzeichnis

Aebi, M. (2003). Spinal metastasis in the elderly. *European Spine Journal*, *12*(SUPPL. 2), S202–S213. https://doi.org/10.1007/s00586-003-0609-9

Algra, P. R., Heimans, J. J., Valk, J., Nauta, J. J., Lachniet, M., & Van Kooten, B. (1992). Do metastases in vertebrae begin in the body or the pedicles? Imaging study in 45 patients. *American Journal of Roentgenology*, *158*(6), 1275–1279. https://doi.org/10.2214/ajr.158.6.1590123

Bartalena, T., Giannelli, G., Rinaldi, M. F., Rimondi, E., Rinaldi, G., Sverzellati, N., & Gavelli, G. (2009). Prevalence of thoracolumbar vertebral fractures on multidetector CT. *European Journal of Radiology*, *69*(3), 555–559. https://doi.org/10.1016/j.ejrad.2007.11.036

Baumgartner, C. F., Koch, L. M., Pollefeys, M., & Konukoglu, E. (2018). An Exploration of 2D and 3D Deep Learning Techniques for Cardiac MR Image Segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 10663 LNCS* (pp. 111–119). https://doi.org/10.1007/978-3-319-75541-0_12

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., West, R. B., van de Rijn, M., & Koller, D. (2011). Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Science Translational Medicine*, *3*(108), 108ra113-108ra113. https://doi.org/10.1126/scitranslmed.3002564

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, *84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Black, D. M., Delmas, P. D., Eastell, R., Reid, I. R., Boonen, S., Cauley, J. A., Cosman, F., Lakatos, P., Leung, P. C., Man, Z., Mautalen, C., Mesenbrink, P., Hu, H., Caminis, J., Tong, K., Rosario-Jansen, T., Krasnow, J., Hue, T. F., Sellmeyer, D., … Cummings, S. R. (2007). Once-Yearly Zoledronic Acid for Treatment of Postmenopausal Osteoporosis. *New England Journal of Medicine*, *356*(18), 1809–1822. https://doi.org/10.1056/NEJMoa067312

Blasco, J., Martinez-Ferrer, A., Macho, J., San Roman, L., Pomés, J., Carrasco, J., Monegal, A., Guañabens, N., & Peris, P. (2012). Effect of vertebroplasty on pain relief, quality of life, and the incidence of new vertebral fractures: A 12-month randomized follow-up, controlled trial. *Journal of Bone and Mineral Research*, *27*(5), 1159–1166. https://doi.org/10.1002/jbmr.1564

Cabitza, F., Cameli, M., Campagner, A., Natali, C., & Ronzio, L. (2022). *Painting the black box white: experimental findings from applying XAI to an ECG reading setting*. 269–286.

Calabrese, E., Rudie, J. D., Rauschecker, A. M., Villanueva-Meyer, J. E., Clarke, J. L., Solomon, D. A., & Cha, S. (2022). Combining radiomics and deep convolutional neural network features from preoperative MRI for predicting clinically relevant genetic biomarkers in glioblastoma. *Neuro-Oncology Advances*, *4*(1), 1–11. https://doi.org/10.1093/noajnl/vdac060

Cataldi, V., Laporta, T., Sverzellati, N., De Filippo, M., & Zompatori, M. (2008). Detection of incidental vertebral fractures on routine lateral chest radiographs. *La Radiologia Medica*, *113*(7), 968–977. https://doi.org/10.1007/s11547-008-0294-1

Cauley, J. A., Hochberg, M. C., Lui, L.-Y., Palermo, L., Ensrud, K. E., Hillier, T. A., Nevitt, M. C., & Cummings, S. R. (2007). Long-term Risk of Incident Vertebral Fractures. *JAMA*, *298*(23), 2761. https://doi.org/10.1001/jama.298.23.2761

Chan, H.-P., Samala, R. K., Hadjiiski, L. M., & Zhou, C. (2020). Deep Learning in Medical Image Analysis. In *Advances in Experimental Medicine and Biology* (Vol. 1213, pp. 3–21). https://doi.org/10.1007/978-3-030-33128-3_1

Chen, S., Sedghi Gamechi, Z., Dubost, F., van Tulder, G., & de Bruijne, M. (2022). An end-to-end approach to segmentation in medical images with CNN and posterior-CRF. *Medical Image Analysis*, *76*, 102311. https://doi.org/10.1016/j.media.2021.102311

Cho, J. H., Ha, J.-K., Hwang, C. J., Lee, D.-H., & Lee, C. S. (2015). Patterns of Treatment for Metastatic Pathological Fractures of the Spine: The Efficacy of Each Treatment Modality. *Clinics in Orthopedic Surgery*, *7*(4), 476. https://doi.org/10.4055/cios.2015.7.4.476

Chong, S., Shin, S.-H., Yoo, H., Lee, S. H., Kim, K.-J., Jahng, T.-A., & Gwak, H.-S. (2012). Single-stage posterior decompression and stabilization for metastasis of the thoracic spine: prognostic factors for functional outcome and patients' survival. *The Spine Journal*, *12*(12), 1083–1092. https://doi.org/10.1016/j.spinee.2012.10.015

Chow, R., Harrison, J., & Dornan, J. (1989). Prevention and rehabilitation of osteoporosis program: exercise and osteoporosis. *International Journal of Rehabilitation Research*, *12*(1), 49–56. https://doi.org/10.1097/00004356-198903000-00005

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 9901 LNCS* (pp. 424–432). https://doi.org/10.1007/978-3-319-46723-8_49

Cooper, C., Atkinson, E. J., Jacobsen, S. J., O'Fallon, W. M., & Melton, L. J. (1993). Population-Based Study of Survival after Osteoporotic Fractures. *American Journal of Epidemiology*, *137*(9), 1001–1005. https://doi.org/10.1093/oxfordjournals.aje.a116756

Cooper, C., O'Neill, T., & Silman, A. (1993). The epidemiology of vertebral fractures. *Bone*, *14*(SUPPL. 1), 89–97. https://doi.org/10.1016/8756-3282(93)90358-H

Cuénod, C. A., Laredo, J. D., Chevret, S., Hamze, B., Naouri, J. F., Chapaux, X., Bondeville, J. M., & Tubiana, J. M. (1996). Acute vertebral collapse due to osteoporosis or malignancy: appearance on unenhanced and gadolinium-enhanced MR images. *Radiology*, *199*(2), 541–549. https://doi.org/10.1148/radiology.199.2.8668809

Cummings, S. R., Martin, J. S., McClung, M. R., Siris, E. S., Eastell, R., Reid, I. R., Delmas, P., Zoog, H. B., Austin, M., Wang, A., Kutilek, S., Adami, S., Zanchetta, J., Libanati, C., Siddhanti, S., & Christiansen, C. (2009). Denosumab for Prevention of Fractures in Postmenopausal Women With Osteoporosis. *Obstetrical &*

Gynecological Survey, *64*(12), 805–807. https://doi.org/10.1097/01.ogx.0000363236.41902.96

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., & Heng, P.-A. (2017). 3D deeply supervised network for automated segmentation of volumetric medical images. *Medical Image Analysis*, *41*, 40–54. https://doi.org/10.1016/j.media.2017.05.001

Ettinger, B., Black, D. M., Nevitt, M. C., Rundle, A. C., Cauley, J. A., Cummings, S. R., & Genant, H. K. (1992). Contribution of vertebral deformities to chronic back pain and disability. *Journal of Bone and Mineral Research*, *7*(4), 449–456. https://doi.org/10.1002/jbmr.5650070413

Fernández, S. S., Miralles, F., Serrato, A. R., Alegría, J. G., Cantero, A. R., Ordoñez, M. A. G., Terán, C. M. S. R., Zorzano, E. G., & Gómez-Huelgas, R. (2012). Prevalence of thoracic vertebral fractures in Spanish patients hospitalized in Internal Medicine Departments. Assessment of the clinical inertia. (PREFRAMI study). *European Journal of Internal Medicine*, *23*(2), e44–e47. https://doi.org/10.1016/j.ejim.2011.11.015

Ferrar, L., Roux, C., Felsenberg, D., Glüer, C.-C., & Eastell, R. (2012). Association between incident and baseline vertebral fractures in European women: vertebral fracture assessment in the Osteoporosis and Ultrasound Study (OPUS). *Osteoporosis International*, *23*(1), 59–65. https://doi.org/10.1007/s00198-011-1701-3

Filograna, L., Lenkowicz, J., Cellini, F., Dinapoli, N., Manfrida, S., Magarelli, N., Leone, A., Colosimo, C., & Valentini, V. (2019). Identification of the most significant magnetic resonance imaging (MRI) radiomic features in oncological patients with vertebral bone marrow metastatic disease: a feasibility study. *La Radiologia Medica*, *124*(1), 50–57. https://doi.org/10.1007/s11547-018-0935-y

Fink, H. A., Ensrud, K. E., Nelson, D. B., Kerani(formerly Pieper), R. P., Schreiner, P. J., Zhao, Y., Cummings, S. R., & Nevitt, M. C. (2003). Disability after clinical fracture in postmenopausal women with low bone density: the fracture intervention trial (FIT). *Osteoporosis International*, *14*(1), 69–76. https://doi.org/10.1007/s00198-002-1314-y

Fink, H. A., Milavetz, D. L., Palermo, L., Nevitt, M. C., Cauley, J. A., Genant, H. K., Black, D. M., & Ensrud, K. E. (2005). What Proportion of Incident Radiographic Vertebral Deformities Is Clinically Diagnosed and Vice Versa? *Journal of Bone and Mineral Research*, *20*(7), 1216–1222. https://doi.org/10.1359/JBMR.050314

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399

Foreman, S., Gersing, A., Kirschke, J., Schinz, D., Husseini, M. E., Dietrich, A.-S., Weissinger, J., Renz, M., Metz, M., Feuerriegel, G., Wiestler, B., Stahl, R., Wörtler, K., Schwaiger, B. J., & Makowski, M. R. (2024). Deep learning to differentiate benign and malignant vertebral fractures on multidetector CT (accepted and in press). *Radiology*.

Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A. P., & Palmer, L. J. (2017). *Detecting hip fractures with radiologist-level performance using deep neural networks*.

Geith, T., Reiser, M., & Baur-Melnyk, A. (2015). Unterscheidung akuter osteoporotischer und metastasenbedingter Wirbelkörperfrakturen in der Bildgebung. *Der Unfallchirurg*, *118*(3), 222–229. https://doi.org/10.1007/s00113-014-2690-4

Gitto, S., Cuocolo, R., Albano, D., Chianca, V., Messina, C., Gambino, A., Ugga, L., Cortese, M. C., Lazzara, A., Ricci, D., Spairani, R., Zanchetta, E., Luzzati, A., Brunetti, A., Parafioriti, A., & Sconfienza, L. M. (2020). MRI radiomics-based machine-learning classification of bone chondrosarcoma. *European Journal of Radiology*, *128*(April), 109043. https://doi.org/10.1016/j.ejrad.2020.109043

Greenspan, S. L., von Stetten, E., Emond, S. K., Jones, L., & Parker, R. A. (2001). Instant Vertebral Assessment. *Journal of Clinical Densitometry*, *4*(4), 373–380. https://doi.org/10.1385/JCD:4:4:373

Grewal, M., Srivastava, M. M., Kumar, P., & Varadarajan, S. (2018). RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, *2018-April*, 281–284. https://doi.org/10.1109/ISBI.2018.8363574

Guo, K., Li, X., Hu, X., Liu, J., & Fan, T. (2021). Hahn-PCNN-CNN: an end-to-end multi-modal brain medical image fusion framework useful for clinical diagnosis. *BMC Medical Imaging*, *21*(1), 111. https://doi.org/10.1186/s12880-021-00642-z

Gyftopoulos, S., Kim, D., Aaltonen, E., & Horwitz, L. I. (2016). Patient Recall Imaging in the Ambulatory Setting. *American Journal of Roentgenology*, *206*(4), 787–791. https://doi.org/10.2214/AJR.15.15268

Ha, K.-Y., Min, C.-K., Seo, J.-Y., Kim, Y.-H., Ahn, J.-H., Hyun, N.-M., & Kim, Y.-C. (2015). Bone Cement Augmentation Procedures for Spinal Pathologic Fractures by Multiple Myeloma. *Journal of Korean Medical Science*, *30*(1), 88. https://doi.org/10.3346/jkms.2015.30.1.88

Harrison, J. E., Chow, R., Dornan, J., Goodwin, S., & Strauss, A. (1993). Evaluation of a program for rehabilitation of osteoporotic patients (PRO): 4-year follow-up. *Osteoporosis International*, *3*(1), 13–17. https://doi.org/10.1007/BF01623171

He, Y., Pan, I., Bao, B., Halsey, K., Chang, M., Liu, H., Peng, S., Sebro, R. A., Guan, J., Yi, T., Delworth, A. T., Eweje, F., States, L. J., Zhang, P. J., Zhang, Z., Wu, J., Peng, X., & Bai, H. X. (2020). Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. *EBioMedicine*, *62*, 103121. https://doi.org/10.1016/j.ebiom.2020.103121

Hesamian, M. H., Jia, W., He, X., & Kennedy, P. (2019). Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging*, *32*(4), 582–596. https://doi.org/10.1007/s10278-019-00227-x

Hirabayashi, H., Ebara, S., Kinoshita, T., Yuzawa, Y., Nakamura, I., Takahashi, J., Kamimura, M., Ohtsuka, K., & Takaoka, K. (2003). Clinical outcome and survival after palliative surgery for spinal metastases. *Cancer*, *97*(2), 476–484. https://doi.org/10.1002/cncr.11039

Hirschmann, A., Cyriac, J., Stieltjes, B., Kober, T., Richiardi, J., & Omoumi, P. (2019). Artificial Intelligence in Musculoskeletal Imaging: Review of Current Literature, Challenges, and Trends. *Seminars in Musculoskeletal Radiology*, *23*(03), 304–311. https://doi.org/10.1055/s-0039-1684024

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. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12266 LNCS* (pp. 733–742). https://doi.org/10.1007/978-3-030-59725-2_71

Ibrahim, A., Crockard, A., Antonietti, P., Boriani, S., Bünger, C., Gasbarrini, A., Grejs, A., Harms, J., Kawahara, N., Mazel, C., Melcher, R., & Tomita, K. (2008). Does spinal surgery improve the quality of life for those with extradural (spinal) osseous metastases? An international multicenter prospective observational study of 223 patients. *Journal of Neurosurgery: Spine*, *8*(3), 271–278. https://doi.org/10.3171/SPI/2008/8/3/271

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *32nd International Conference on Machine Learning, ICML 2015*, *1*, 448–456.

Jackson, S. A., Tenenhouse, A., & Robertson, L. (2000). Vertebral Fracture Definition from Population-Based Data: Preliminary Results from the Canadian Multicenter Osteoporosis Study (CaMos). *Osteoporosis International*, *11*(8), 680–687. https://doi.org/10.1007/s001980070066

Jamaludin, A., Lootus, M., Kadir, T., Zisserman, A., Urban, J., Battié, M. C., Fairbank, J., & McCall, I. (2017). ISSLS PRIZE IN BIOENGINEERING SCIENCE 2017: Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. *European Spine Journal*, *26*(5), 1374–1383. https://doi.org/10.1007/s00586-017-4956-3

Jha, S., & Topol, E. J. (2016). Adapting to Artificial Intelligence. *JAMA*, *316*(22), 2353. https://doi.org/10.1001/jama.2016.17438

Johnell, O., Kanis, J. A., Odén, A., Sernbo, I., Redlund-Johnell, I., Petterson, C., De Laet, C., & Jönsson, B. (2004). Fracture risk following an osteoporotic fracture. *Osteoporosis International*, *15*(3), 175–179. https://doi.org/10.1007/s00198-003-1514-0

Jung, H.-S., Jee, W.-H., McCauley, T. R., Ha, K.-Y., & Choi, K.-H. (2003). Discrimination of Metastatic from Acute Osteoporotic Compression Spinal Fractures with MR Imaging 1. *RadioGraphics*, *23*(1), 179–187. https://doi.org/10.1148/rg.231025043

Kalmet, P. H. S., Sanduleanu, S., Primakov, S., Wu, G., Jochems, A., Refaee, T., Ibrahim, A., Hulst, L. v., Lambin, P., & Poeze, M. (2020). Deep learning in fracture detection: a narrative review. *Acta Orthopaedica*, *91*(2), 215–220. https://doi.org/10.1080/17453674.2019.1711323

Karchevsky, M., Babb, J. S., & Schweitzer, M. E. (2008). Can diffusion-weighted imaging be used to differentiate benign from pathologic fractures? A meta-analysis. *Skeletal Radiology*, *37*(9), 791–795. https://doi.org/10.1007/s00256-008-0503-y

Kawahara, J., BenTaieb, A., & Hamarneh, G. (2016). Deep features to classify skin lesions. *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, *2016-June*, 1397–1400. https://doi.org/10.1109/ISBI.2016.7493528

Kendler, D. L., Bauer, D. C., Davison, K. S., Dian, L., Hanley, D. A., Harris, S. T., McClung, M. R., Miller, P. D., Schousboe, J. T., Yuen, C. K., & Lewiecki, E. M.

(2016). Vertebral Fractures: Clinical Importance and Management. *The American Journal of Medicine*, *129*(2), 221.e1-221.e10. https://doi.org/10.1016/j.amjmed.2015.09.020

Kim, D. H., & MacKinnon, T. (2018). Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology*, *73*(5), 439–445. https://doi.org/10.1016/j.crad.2017.11.015

Kim, N., Rowe, B. H., Raymond, G., Jen, H., Colman, I., Jackson, S. A., Siminoski, K. G., Chahal, A. M., Folk, D., & Majumdar, S. R. (2004). Underreporting of Vertebral Fractures on Routine Chest Radiography. *American Journal of Roentgenology*, *182*(2), 297–300. https://doi.org/10.2214/ajr.182.2.1820297

Kim, W. H., Hur, G., Woo, J. J., Cho, W. H., & Jung, M. J. (1995). MRI of Vertebral Compression Fractures: Differentiation between Benign and Malignant Causes. *Journal of the Korean Radiological Society*, *33*(5), 673. https://doi.org/10.3348/jkrs.1995.33.5.673

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, *35*, 303–312. https://doi.org/10.1016/j.media.2016.07.007

Kwon, Y. M., Kim, K. S., Kuh, S. U., Chin, D. K., Jin, B. H., & Cho, Y. E. (2009). Survival Rate and Neurological Outcome after Operation for Advanced Spinal Metastasis (Tomita's Classification ≥ Type 4). *Yonsei Medical Journal*, *50*(5), 689. https://doi.org/10.3349/ymj.2009.50.5.689

Langs, G., Attenberger, U., Licandro, R., Hofmanninger, J., Perkonigg, M., Zusag, M., Röhrich, S., Sobotka, D., & Prosch, H. (2020). Maschinelles Lernen in der Radiologie. *Der Radiologe*, *60*(1), 6–14. https://doi.org/10.1007/s00117-019-00624-x

Langs, G., Röhrich, S., Hofmanninger, J., Prayer, F., Pan, J., Herold, C., & Prosch, H. (2018). Machine learning: from radiomics to discovery and routine. *Der Radiologe*, *58*(S1), 1–6. https://doi.org/10.1007/s00117-018-0407-3

Laredo, J. D., Lakhdari, K., Bellaïche, L., Hamze, B., Janklewicz, P., & Tubiana, J. M. (1995). Acute vertebral collapse: CT findings in benign and malignant nontraumatic cases. *Radiology*, *194*(1), 41–48. https://doi.org/10.1148/radiology.194.1.7997579

Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V., & Langlotz, C. P. (2018). Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. *Radiology*, *287*(1), 313–322. https://doi.org/10.1148/radiol.2017170236

Lee, Y. H. (2018). Efficiency Improvement in a Busy Radiology Practice: Determination of Musculoskeletal Magnetic Resonance Imaging Protocol Using Deep-Learning Convolutional Neural Networks. *Journal of Digital Imaging*, *31*(5), 604–610. https://doi.org/10.1007/s10278-018-0066-y

Lentle, B., Koromani, F., Brown, J. P., Oei, L., Ward, L., Goltzman, D., Rivadeneira, F., Leslie, W. D., Probyn, L., Prior, J., Hammond, I., Cheung, A. M., & Oei, E. H. (2019). The Radiology of Osteoporotic Vertebral Fractures Revisited. *Journal of Bone and Mineral Research*, *34*(3), 409–418. https://doi.org/10.1002/jbmr.3669

Li, Y., Zhang, Y., Zhang, E., Chen, Y., Wang, Q., Liu, K., Yu, H. J., Yuan, H., Lang, N., & Su, M.-Y. (2021). Differential diagnosis of benign and malignant vertebral fracture

on CT using deep learning. *European Radiology*, *31*(12), 9612–9619. https://doi.org/10.1007/s00330-021-08014-5
Liew, C. (2018). The future of radiology augmented with Artificial Intelligence: A strategy for success. *European Journal of Radiology*, *102*(December 2017), 152–156. https://doi.org/10.1016/j.ejrad.2018.03.019
Lim, B.-S., Chang, U.-K., & Youn, S.-M. (2009). Clinical Outcomes after Percutaneous Vertebroplasty for Pathologic Compression Fractures in Osteolytic Metastatic Spinal Disease. *Journal of Korean Neurosurgical Society*, *45*(6), 369. https://doi.org/10.3340/jkns.2009.45.6.369
Liu, R., Pan, D., Xu, Y., Zeng, H., He, Z., Lin, J., Zeng, W., Wu, Z., Luo, Z., Qin, G., & Chen, W. (2022). A deep learning–machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors. *European Radiology*, *32*(2), 1371–1383. https://doi.org/10.1007/s00330-021-08195-z
Löffler, M. T., Sekuboyina, A., Jacob, A., Grau, A., Scharr, A., El Husseini, M., Kallweit, M., Zimmer, C., Baum, T., & Kirschke, J. S. (2020). A Vertebral Segmentation Dataset with Fracture Grading. *Radiology: Artificial Intelligence*, *2*(4), e190138. https://doi.org/10.1148/ryai.2020190138
Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift Für Medizinische Physik*, *29*(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002
MacLean, C., Newberry, S., Maglione, M., McMahon, M., Ranganath, V., Suttorp, M., Mojica, W., Timmer, M., Alexander, A., McNamara, M., Desai, S. B., Zhou, A., Chen, S., Carter, J., Tringale, C., Valentine, D., Johnsen, B., & Grossman, J. (2008). Systematic Review: Comparative Effectiveness of Treatments to Prevent Fractures in Men and Women with Low Bone Density or Osteoporosis. *Annals of Internal Medicine*, *148*(3), 197. https://doi.org/10.7326/0003-4819-148-3-200802050-00198
Majumdar, S. R., Kim, N., Colman, I., Chahal, A. M., Raymond, G., Jen, H., Siminoski, K. G., Hanley, D. A., & Rowe, B. H. (2005). Incidental Vertebral Fractures Discovered With Chest Radiography in the Emergency Department. *Archives of Internal Medicine*, *165*(8), 905. https://doi.org/10.1001/archinte.165.8.905
Matsumura, R., Harada, K., Domae, Y., & Wan, W. (2019). Learning Based Industrial Bin-Picking Trained with Approximate Physics Simulator. In *Advances in Intelligent Systems and Computing* (Vol. 867, pp. 786–798). https://doi.org/10.1007/978-3-030-01370-7_61
Mauch, J. T., Carr, C. M., Cloft, H., & Diehn, F. E. (2018). Review of the Imaging Features of Benign Osteoporotic and Malignant Vertebral Compression Fractures. *American Journal of Neuroradiology*, *39*(9), 1584–1592. https://doi.org/10.3174/ajnr.A5528
Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2019). Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of Magnetic Resonance Imaging*, *49*(4), 939–954. https://doi.org/10.1002/jmri.26534
McBee, M. P., Awan, O. A., Colucci, A. T., Ghobadi, C. W., Kadom, N., Kansagra, A. P., Tridandapani, S., & Auffermann, W. F. (2018). Deep Learning in Radiology.

Academic Radiology, *25*(11), 1472–1480. https://doi.org/10.1016/j.acra.2018.02.018
Melton, J. L. (1997). Epidemiology of Spinal Osteoporosis. *Spine*, *22*(Supplement), 2S-11S. https://doi.org/10.1097/00007632-199712151-00002
Melton, L. J., Lane, A. W., Cooper, C., Eastell, R., O'Fallon, W. M., & Riggs, B. L. (1993). Prevalence and incidence of vertebral deformities. *Osteoporosis International*, *3*(3), 113–119. https://doi.org/10.1007/BF01623271
Merkow, J., Lufkin, R., Nguyen, K., Soatto, S., Tu, Z., & Vedaldi, A. (2017). *DeepRadiologyNet: Radiologist Level Pathology Detection in CT Head Images*. 1–22.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236
Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., & Bowling, M. (2017). DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, *356*(6337), 508–513. https://doi.org/10.1126/science.aam6960
Moulopoulos, L. A., Yoshimitsu, K., Johnston, D. A., Leeds, N. E., & Libshitz, H. I. (1996). MR prediction of benign and malignant vertebral compression fractures. *Journal of Magnetic Resonance Imaging*, *6*(4), 667–674. https://doi.org/10.1002/jmri.1880060416
Mudano, A. S., Bian, J., Cope, J. U., Curtis, J. R., Gross, T. P., Allison, J. J., Kim, Y., Briggs, D., Melton, M. E., Xi, J., & Saag, K. G. (2009). Vertebroplasty and kyphoplasty are associated with an increased risk of secondary vertebral compression fractures: a population-based cohort study. *Osteoporosis International*, *20*(5), 819–826. https://doi.org/10.1007/s00198-008-0745-5
Mui, L. W., Haramati, L. B., Alterman, D. D., Haramati, N., Zelefsky, M. N., & Hamerman, D. (2003). Evaluation of Vertebral Fractures on Lateral Chest Radiographs of Inner-City Postmenopausal Women. *Calcified Tissue International*, *73*(6), 550–554. https://doi.org/10.1007/s00223-003-0064-y
Murata, K., Endo, K., Aihara, T., Suzuki, H., Sawaji, Y., Matsuoka, Y., Nishimura, H., Takamatsu, T., Konishi, T., Maekawa, A., Yamauchi, H., Kanazawa, K., Endo, H., Tsuji, H., Inoue, S., Fukushima, N., Kikuchi, H., Sato, H., & Yamamoto, K. (2020). Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. *Scientific Reports*, *10*(1), 20031. https://doi.org/10.1038/s41598-020-76866-w
Navarro, F., Dapper, H., Asadpour, R., Knebel, C., Spraker, M. B., Schwarze, V., Schaub, S. K., Mayr, N. A., Specht, K., Woodruff, H. C., Lambin, P., Gersing, A. S., Nyflot, M. J., Menze, B. H., Combs, S. E., & Peeken, J. C. (2021). Development and External Validation of Deep-Learning-Based Tumor Grading Models in Soft-Tissue Sarcoma Patients Using MR Imaging. *Cancers*, *13*(12), 2866. https://doi.org/10.3390/cancers13122866

Nevitt, M. C. (1998). The Association of Radiographically Detected Vertebral Fractures with Back Pain and Function: A Prospective Study. *Annals of Internal Medicine*, *128*(10), 793. https://doi.org/10.7326/0003-4819-128-10-199805150-00001
Obaid, H., Husamaldin, Z., & Bhatt, R. (2008). Underdiagnosis of vertebral collapse on routine multidetector computed tomography scan of the abdomen. *Acta Radiologica*, *49*(7), 795–800. https://doi.org/10.1080/02841850802165776
Olczak, J., Fahlberg, N., Maki, A., Razavian, A. S., Jilert, A., Stark, A., Sköldenberg, O., & Gordon, M. (2017). Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthopaedica*, *88*(6), 581–586. https://doi.org/10.1080/17453674.2017.1344459
Papaioannou, A., Morin, S., Cheung, A. M., Atkinson, S., Brown, J. P., Feldman, S., Hanley, D. A., Hodsman, A., Jamal, S. A., Kaiser, S. M., Kvern, B., Siminoski, K., & Leslie, W. D. (2010). 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. *Canadian Medical Association Journal*, *182*(17), 1864–1873. https://doi.org/10.1503/cmaj.100771
Park, T., Yoon, M. A., Cho, Y. C., Ham, S. J., Ko, Y., Kim, S., Jeong, H., & Lee, J. (2022). Publisher Correction: Automated segmentation of the fractured vertebrae on CT and its applicability in a radiomics model to predict fracture malignancy. *Scientific Reports*, *12*(1), 7171. https://doi.org/10.1038/s41598-022-11784-7
Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. 3–9.
Recht, M., & Bryan, R. N. (2017). Artificial Intelligence: Threat or Boon to Radiologists? *Journal of the American College of Radiology*, *14*(11), 1476–1480. https://doi.org/10.1016/j.jacr.2017.07.007
Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., Summers, R. M., & Giger, M. L. (2019). Deep learning in medical imaging and radiation therapy. *Medical Physics*, *46*(1), e1–e36. https://doi.org/10.1002/mp.13264
Sato, Y., Takegami, Y., Asamoto, T., Ono, Y., Hidetoshi, T., Goto, R., Kitamura, A., & Honda, S. (2021). Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: a multicenter study. *BMC Musculoskeletal Disorders*, *22*(1), 407. https://doi.org/10.1186/s12891-021-04260-2
Sayed-Noor, A. S., Ågren, P.-H., & Wretenberg, P. (2011). Interobserver Reliability and Intraobserver Reproducibility of Three Radiological Classification Systems for Intra-articular Calcaneal Fractures. *Foot & Ankle International*, *32*(9), 861–866. https://doi.org/10.3113/FAI.2011.0861
Schousboe, J. T., Ensrud, K. E., Nyman, J. A., Kane, R. L., & Melton, L. J. (2006). Cost-Effectiveness of Vertebral Fracture Assessment to Detect Prevalent Vertebral Deformity and Select Postmenopausal Women With a Femoral Neck T-Score > -2.5 for Alendronate Therapy: A Modeling Study. *Journal of Clinical Densitometry*, *9*(2), 133–143. https://doi.org/10.1016/j.jocd.2005.11.004
Schuster, J. M., & Grady, M. S. (2001). Medical management and adjuvant therapies in spinal metastatic disease. *Neurosurgical Focus*, *11*(6), 1–3. https://doi.org/10.3171/foc.2001.11.6.4
Schwaiger, B., Gersing, A., Baum, T., Krestan, C., & Kirschke, J. (2016). Distinguishing Benign and Malignant Vertebral Fractures Using CT and MRI. *Seminars in*

Musculoskeletal Radiology, *20*(04), 345–352. https://doi.org/10.1055/s-0036-1592433
Sekuboyina, A., Husseini, M. E., Bayat, A., Löffler, M., Liebl, H., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., Urschler, M., Chen, M., Cheng, D., Lessmann, N., Hu, Y., Wang, T., Yang, D., Xu, D., Ambellan, F., … Kirschke, J. S. (2021). VerSe: A Vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Analysis*, *73*, 102166. https://doi.org/10.1016/j.media.2021.102166
Sekuboyina, A., Rempfler, M., Valentinitsch, A., Menze, B. H., & Kirschke, J. S. (2020). Labeling Vertebrae with Two-dimensional Reformations of Multidetector CT Images: An Adversarial Approach for Incorporating Prior Knowledge of Spine Anatomy. *Radiology: Artificial Intelligence*, *2*(2), e190074. https://doi.org/10.1148/ryai.2020190074
Shehovych, A., Salar, O., Meyer, C., & Ford, D. (2016). Adult distal radius fractures classification systems: essential clinical knowledge or abstract memory testing? *The Annals of The Royal College of Surgeons of England*, *98*(8), 525–531. https://doi.org/10.1308/rcsann.2016.0237
Shon, I. H., & Fogelman, I. (2003). F-18 FDG Positron Emission Tomography and Benign Fractures. *Clinical Nuclear Medicine*, *28*(3), 171–175. https://doi.org/10.1097/01.RLU.0000053508.98025.01
Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–14.
Sinaki, M. (2012). Exercise for Patients With Osteoporosis: Management of Vertebral Compression Fractures and Trunk Strengthening for Fall Prevention. *PM&R*, *4*(11), 882–888. https://doi.org/10.1016/j.pmrj.2012.10.008
Sinaki, M., Itoi, E., Wahner, H. W., Wollan, P., Gelzcer, R., Mullan, B. P., Collins, D. A., & Hodgson, S. F. (2002). Stronger back muscles reduce the incidence of vertebral fractures: a prospective 10 year follow-up of postmenopausal women. *Bone*, *30*(6), 836–841. https://doi.org/10.1016/S8756-3282(02)00739-1
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.
Sun, Q., Lin, X., Zhao, Y., Li, L., Yan, K., Liang, D., Sun, D., & Li, Z. (2020). Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region. *Frontiers in Oncology*, *10*(January), 1–12. https://doi.org/10.3389/fonc.2020.00053
The Board of Trustees of The North American Menopause Society. (2010). Management of osteoporosis in postmenopausal women. *Menopause*, *17*(1), 25–54. https://doi.org/10.1097/gme.0b013e3181c617e6
The Cancer Imaging Archive. (2023). https://www.cancerimagingarchive.net
Tomita, N., Cheung, Y. Y., & Hassanpour, S. (2018). Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Computers in Biology and Medicine*, *98*, 8–15. https://doi.org/10.1016/j.compbiomed.2018.05.011
Torres, C., & Hammond, I. (2016). Computed Tomography and Magnetic Resonance Imaging in the Differentiation of Osteoporotic Fractures From Neoplastic Metastatic

Fractures. *Journal of Clinical Densitometry*, *19*(1), 63–69. https://doi.org/10.1016/j.jocd.2015.08.008
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
van Geel, T. A. C. M., Huntjens, K. M. B., van den Bergh, J. P. W., Dinant, G.-J., & Geusens, P. P. (2010). Timing of Subsequent Fractures after an Initial Fracture. *Current Osteoporosis Reports*, *8*(3), 118–122. https://doi.org/10.1007/s11914-010-0023-2
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
von Schacky, C. E., Wilhelm, N. J., Schäfer, V. S., Leonhardt, Y., Jung, M., Jungmann, P. M., Russe, M. F., Foreman, S. C., Gassert, F. G., Gassert, F. T., Schwaiger, B. J., Mogler, C., Knebel, C., von Eisenhart-Rothe, R., Makowski, M. R., Woertler, K., Burgkart, R., & Gersing, A. S. (2022). Development and evaluation of machine learning models based on X-ray radiomics for the classification and differentiation of malignant and benign bone tumors. *European Radiology*, *32*(9), 6247–6257. https://doi.org/10.1007/s00330-022-08764-w
Waterloo, S., Ahmed, L. A., Center, J. R., Eisman, J. A., Morseth, B., Nguyen, N. D., Nguyen, T., Sogaard, A. J., & Emaus, N. (2012). Prevalence of vertebral fractures in women and men in the population-based Tromsø Study. *BMC Musculoskeletal Disorders*, *13*(1), 3. https://doi.org/10.1186/1471-2474-13-3
Williams, A. L., Al-Busaidi, A., Sparrow, P. J., Adams, J. E., & Whitehouse, R. W. (2009). Under-reporting of osteoporotic vertebral fractures on computed tomography. *European Journal of Radiology*, *69*(1), 179–183. https://doi.org/10.1016/j.ejrad.2007.08.028
Woo, E. K., Mansoubi, H., & Alyas, F. (2008). Incidental vertebral fractures on multidetector CT images of the chest: prevalence and recognition. *Clinical Radiology*, *63*(2), 160–164. https://doi.org/10.1016/j.crad.2007.01.031
Xu, F., Xiong, Y., Ye, G., Liang, Y., Guo, W., Deng, Q., Wu, L., Jia, W., Wu, D., Chen, S., Liang, Z., & Zeng, X. (2023). Deep learning-based artificial intelligence model for classification of vertebral compression fractures: A multicenter diagnostic study. *Frontiers in Endocrinology*, *14*, 1025749. https://doi.org/10.3389/fendo.2023.1025749
Yamashita, T., Aota, Y., Kushida, K., Murayama, H., Hiruma, T., Takeyama, M., Iwamura, Y., & Saito, T. (2008). Changes in Physical Function After Palliative Surgery for Metastatic Spinal Tumor. *Spine*, *33*(21), 2341–2346. https://doi.org/10.1097/BRS.0b013e3181878733
York, T., Jenney, H., & Jones, G. (2020). Clinician and computer: a study on patient perceptions of artificial intelligence in skeletal radiography. *BMJ Health & Care Informatics*, *27*(3), e100233. https://doi.org/10.1136/bmjhci-2020-100233

Yu, L., Chen, H., Dou, Q., Qin, J., & Heng, P.-A. (2017). Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Transactions on Medical Imaging*, *36*(4), 994–1004. https://doi.org/10.1109/TMI.2016.2642839

Zaharchuk, G., Gong, E., Wintermark, M., Rubin, D., & Langlotz, C. P. (2018). Deep Learning in Neuroradiology. *American Journal of Neuroradiology*, *39*(10), 1776–1784. https://doi.org/10.3174/ajnr.A5543

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss Functions for Image Restoration With Neural Networks. *IEEE Transactions on Computational Imaging*, *3*(1), 47–57. https://doi.org/10.1109/TCI.2016.2644865

Abbildungs-und Tabellenverzeichnis

Danksagung

An dieser Stelle möchte ich allen beteiligten Personen, die mich bei der Anfertigung meiner Dissertation unterstützt haben, meinen großen Dank aussprechen.

Mein besonderer Dank gilt Prof. Dr. med Alexandra Gersing für die hervorragende Betreuung und das hilfreiche Feedback.

Außerdem möchte ich mich bei PD Dr. med Benedikt J. Schwaiger für seine Arbeit als Mentor bedanken.

Besonders danken möchte ich PD Dr. med Sarah Foreman für die hervorragende Einarbeitung und die exzellente Betreuung.

Des Weiteren will ich mich bei der gesamten muskuloskelettalen Forschungsgruppe des Klinikums Rechts der Isar bedanken.

Bei dieser Gelegenheit will ich zudem der Studienstiftung des Deutsches Volkes für die Unterstützung während des gesamten Studiums meinen herzlichen Dank aussprechen.

Ganz herzlich will ich mich auch bei meiner Familie, Dr. med. Maria Dietrich, Dr. med. Wolfgang Dietrich und Julia Dietrich, sowie bei Christine und Andreas Weiß und Editha Dietrich und Dr. med. vet. Wolfgang Dietrich, bedanken.