Technische Universität München
Fakultät für Medizin

Evaluation der diagnostischen Genauigkeit von Machine Learning für die Detektion von Lungenrundherden in der Thoraxradiographie

Marie Lodde

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades einer Doktorin der Medizin genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Marcus Makowski
Prüfer*innen der Dissertation:	1. Prof. Dr. Daniela Pfeiffer
	2. apl. Prof. Dr. Jan Stefan Kirschke

Die Dissertation wurde am 24.02.2022 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 07.06.2022 angenommen.

I Inhaltsverzeichnis

II Abkürzungsverzeichnis

Abkürzung	Beschreibung
Abb.	Abbildung
AI	Artificial Intelligence
Abk.	Abkürzung
Bspw.	Beispielsweise
Bzgl.	Bezüglich
Bzw.	Beziehungsweise
Ca.	Circa
CADe	Computer-aided detection
CADx	Computer-aided diagnosis
Cm	Centimeter
CNN	Convolutional Neural Network
CT	Computertomographie
DICOM	Digital Imaging and Communications in Medicine
D.h.	Das heißt
DIMDI	Deutsches Institut für medizinische Dokumentation und Information
Dt.	Deutsch
EKG	Elektrokardiogramm
Engl.	Englisch
Etc.	Et cetera
FN	False Negative/Falsch Negativ
FP	False Positive/Falsch Positiv
FPR	False Positive Rate/Falsch Positiv Rate
Ggf.	Gegebenenfalls
Kap.	Kapitel

KI	Künstliche Intelligenz
ICD	International Statistical Classification of Disease and Related Health Problems
ICD-10-GM	International Statistical Classification of Diseases and Related Health Problems-10-German Modification - 10. Revision der Klassifikation
IT-System	Informationstechnisches System
JSRT	Japanese Society of Radiological Technology
LRH	Lungenrundherd
MSv	Millisievert
Od.	Oder
O. n. A.	Ohne nähere Angabe(n)
PACS	Picture Archiving and Communication System
S.	Seite
S.g.	So genannt
SPN	Solitaire Pulmonal Nodule
RH	Rundherd
RH +	Rundherd und Nebenerkrankung
RH + FM	Rundherd und Fremdmaterial, kann zusätzlich Nebenerkrankung enthalten
RIS	Radiologie-Informationssystem
TN	True Negative/Richtig Negativ
TNR	True Negative Rate/Richtig Negativ Rate
TP	True Positive/Richtig Positiv
TPR	True Positive Rate/Richtig Positiv Rate
u. ä.	Und ähnlich
U	Unauffällig
U + M	Unauffällig und Mamillenschatten
U +	Unauffällig und Nebenerkrankung

U + FM	Unauffällig und Fremdmaterial, kann zusätzlich Nebenerkrankung enthalten
Vgl.	Vergleiche
Z. B.	Zum Beispiel
Z. T.	Zum Teil

III Abbildungsverzeichnis

IV Tabellenverzeichnis

1 Einführung

Zu Beginn werden in Kapitel 1.1 die Motivation und die Zielsetzung der vorliegenden Arbeit erläutert. Anschließend wird in Abschnitt 1.2 die Herangehensweise an diese Arbeit anhand des Aufbaus dargelegt.

1.1 Motivation und Zielsetzung

Wenn man Ärzte fragt, was die wichtigsten Faktoren für eine erfolgreiche Patientenversorgung sind, fallen häufig zwei Worte: Wissen und Erfahrung. Je mehr ein Arzt weiß und je mehr Patienten er behandelt hat, desto besser kann eine hohe Qualität der Patientenversorgung gewährleistet werden. Das bedeutet, dass Ärzte ihren Erfahrungsschatz durch die Behandlung von Patienten ausweiten und gleichzeitig ihr Wissen über Weiterbildungen vergrößern müssen. Beide Faktoren sind häufig das Produkt fortschreitender Karrieren und somit abhängig von der als Arzt praktizierenden Zeit. (Mintz, 2019)

Die Hauptbeschränkung des menschlichen Geistes bei der Erfassung großer Datenmengen ist in erster Linie die Zeit. In ca. 40 Arbeitsjahren wird ein Radiologe im Schnitt etwa 225.000 MRT/CT-Untersuchungen beurteilen. (Yokota, Goto, Bamba, Kiba, & Yamada, 2017) Im Vergleich dazu wird beim maschinellen Lernen diese Größenordnung als Grundbasis verstanden, welche innerhalb kurzer Zeit um Millionen weitere, analysierte Fälle erweitert werden kann. Neben der Zeit- und Kosteneffizienz, der menschlichen Ermüdbarkeit und Erfahrung, als auch der Objektivierbarkeit von Diagnosestellungen, ist dieses ein essentieller Grund, warum die künstliche Intelligenz (KI) und ihre Fähigkeiten in der Medizin in der Vordergrund gerückt werden. (Mintz, 2019)

Lungenkrebs ist eine Erkrankung, die durch das Auftreten und die unkontrollierte Proliferation von entarteten Lungenzellen gekennzeichnet ist. Diese Krankheit ist mit ca. 1,76 Millionen Todesfällen pro Jahr eine der Hauptursachen für die Sterblichkeit weltweit. (Ferlay et al., 2019) Noch häufiger jedoch sind Lungenmetastasen anderer Primärtumore, wie bspw. aus Bereichen der Mamma, Niere, Kopf, Hals, sowie Gastrointestinaltrakt und der Haut. (Adler, 2011)

Die derzeitige Technologie für eine genaue Diagnose und Auswahl von Behandlungsoptionen basiert auf molekularen Biomarkern, die bei Lungenbiopsien und/oder Blutuntersuchungen eingesetzt werden. Dieser Ansatz kann den Lungenkrebstyp, -stadium und -mutation diagnostizieren. Dennoch bringt dieses Verfahren zur Diagnosesicherung auch einige negative Aspekte mit sich. Patienten müssen sich zu Beginn einem invasiven chirurgischen Eingriff zur Durchführung von Gewebebiopsien unterziehen und eine bestimmte Zeit auf die Identifikation des Krebstyps warten, um im Anschluss daran eine effektive Krebstherapie beginnen zu können. Lange Wartezeiten für diagnostische Prozesse bei Krebserkrankungen sind entscheidende Faktoren, die

das Überleben und die Heilungschancen beeinflussen. Zugleich verursachen Krankenhäuser mit dieser Art von Diagnosestellung hohe Kosten. Zu nennen sind Krankenhausaufenthalte aufgrund invasiver Biopsien, Anschaffungen und Instandhaltungen geeigneter Geräte und Materialien sowie die Beschäftigung spezialisierten Personals. Neue Technologien basieren daher auf künstlicher Intelligenz und Deep-Learning-Techniken, die zur Klassifizierung verschiedener Tumore verwendet werden und somit negative Aspekte bisheriger Diagnosestellungen reduzieren können. (Khosravi, Kazemi, Imielinski, Elemento, & Hajirasouliha, 2017) Die Detektion in den Anfangsstadien gilt als die wirksamste Möglichkeit zur Verbesserung der Überlebenschance eines Patienten. Wird die Pathologie in einem Anfangsstadium nachgewiesen, so beträgt die 5-Jahres-Überlebensrate etwa 54 %. Erfolgt die Detektion in einem fortgeschrittenen Stadium, beträgt die Überlebensrate für 5 Jahre nur 4 %. (Howlader et al., 2014)

Derzeit gilt die Computertomographie als bildgebende Modalität, die am besten für Untersuchungen zur Früherkennung von Lungenkrebs geeignet ist (Firmino, Angelo, Morais, Dantas, & Valentim, 2016). Dennoch ist diese im Vergleich zum konventionellem Röntgen mit einer erhöhten Strahlenbelastung, höheren Kosten (Alkadhi, Frauenfelder, & Schmidt, 2012) und einer geringeren Verfügbarkeit vergesellschaftet (Ngoya, Muhogora, & Pitcher, 2016b).

NGOYA‘S Ergebnisse zeigen eine intuitive, preisgesteuerte Hierarchie des Zugangs zur Bildgebung, wobei die Verfügbarkeit von Modalitäten umgekehrt zu den relativen Kosten in Beziehung steht. Somit ist die kostengünstigste Modalität am meisten verfügbar, und steigende Kosten gehen mit einer sinkenden Verfügbarkeit der Modalitäten einher. Allgemeine Radiographie und Fluoroskopie sind in fast allen geografischen Zonen, so auch zum Beispiel in afrikanischen Ländern mit geringem Einkommen verfügbar und weisen eine homogene Verteilung auf. (Ngoya et al., 2016b) In diesem Zusammenhang erlangten vor allem konventionelle Röntgenaufnahmen des Thorax an klinischer Bedeutung. Aufgrund ihrer weitreichenden Verfügbarkeit ergeben sie eine gute Basis für die Entwicklung von Algorithmen des maschinellen Lernens. Diese Algorithmen wiederum können Radiologen bei der Analyse von Thoraxröntgenbildern unterstützen. (Van Ginneken, 2001)

Basierend auf den zuvor beschriebenen Grundlagen und Fakten soll im Rahmen der vorliegenden Arbeit folgende Forschungsfrage analysiert und im Rahmen einer Reader Studie überprüft werden. Inwieweit können Algorithmen des maschinellen Lernens und praktizierende Radiologen vergleichbare Ergebnisse hinsichtlich der Detektion von Lungenrundherden erzielen und können Algorithmen bereits Anwendung im klinischen Alltag finden? Aufgrund des demographischen Wandels in der Bevölkerung und der damit verbundene Anstieg an Komorbiditäten stellt sich darüber hinaus die Frage, inwiefern die Genauigkeit der Detektion von Lungenrundherden mittels Algorithmen des maschinellen Lernens beim Vorliegen von verschiedenen Fremdmaterialien

beeinträchtigt wird. Auch dieser Fall soll im Rahmen der vorliegenden Arbeit betrachtet werden.

1.2 Aufbau der Arbeit

Die vorliegende Arbeit gliedert sich in sieben Kapitel. Das erste Kapitel umfasst die Einführung, in der die Motivation und die Zielsetzung der Doktorarbeit erläutert werden.
Kapitel zwei umfasst die begrifflichen und konzeptionellen Grundlagen dieser Arbeit. Da eine formale Definition Grundlage einer objektiven Diagnosestellung ist werden die konstituierenden Merkmale und der regelhafte Ablauf eines diagnostischen Prozesses in diesem Zusammenhang genauer definiert. Anschließend werden die Begriffe Lungenrundherd, künstliche Intelligenz und maschinelles Lernen, als auch die Grundlagen von Computer-aided-diagnosis (CADx) und Computer-aided-detection (CADe) erläutert. Den Abschluss dieses Kapitels bildet die Beschreibung der grundlegenden Parameter, die zur Bewertung und zum Vergleich der Ergebnisse dienen.
Im dritten Kapitel werden die bestehenden Ansätze der Forschung genauer untersucht und die Relevanz der vorliegenden Arbeit herausgearbeitet.
Die in dieser Arbeit entwickelte Methodik wird im vierten Kapitel beschrieben. Hierzu werden die technischen Voraussetzungen dargelegt, Anforderungen an die Auswahl der Thoraxröntgenbilder definiert, sowie der Prozess zur Klassifizierung dieser für die verschiedenen Datensätze erläutert. Anschließend wird die Aufbereitung des Datensatzes inklusive der Verwendung der Datensätze, die zum Training, der Validierung und Testung der Algorithmen dienen, erläutert. Diese Datensätze, bestehend aus realen Patientendaten, stellen den klinischen Alltag und die Schwierigkeiten der Diagnostik von Thoraxröntgenbildern wahrheitsgetreu dar. Anschließend wird das Vorgehen zur Analyse der diagnostischen Genauigkeit beschrieben und die Ergebnisdarstellung mit Parametern zur Bewertung der Ergebnisse erläutert.
Im fünften Kapitel erfolgt sodann die Darstellung der Ergebnisse. Kapitel 5.1 zeigt die finale Zusammensetzung der Bilddatensätze. Darauf folgen die Ergebnisse der radiologischen Befundung, der Reader Studie, als auch die Ergebnisse der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial in Kapitel 5.3.
In Kapitel 6 werden die Methoden und die zuvor dargelegten Ergebnisse diskutiert.
Die vorliegende Arbeit schließt mit einer Zusammenfassung der erlangten Kenntnisse im siebten Kapitel ab und gibt einen Ausblick über den weiteren Forschungsbedarf im Themenfeld der Detektion von Lungenrundherden anhand von maschinellem Lernen.

2 Begriffliche und konzeptionelle Grundlagen

Die Entwicklung einer Methodik zur Ermittlung einer objektiven Diagnosestellung bedarf einer einheitlichen Klassifizierung diagnostischer Vorgehensweisen und klarer Begriffsdefinitionen. Die an dieser Stelle der Arbeit vorgenommenen Ausführungen sind die Basis für ein einheitliches Begriffsverständnis. Nachfolgend werden in Kapitel 2.1 die Grundlagen der Diagnostik und in Kapitel 2.2 die Definition und Klassifikation eines Lungenrundherdes beschrieben. Anschließend werden in Kapitel 2.3 und 2.4 künstliche Intelligenz sowie maschinelles Lernen definiert und ihre Funktionen und Eigenschaften erläutert.

2.1 Diagnostik

Diagnostik, abgeleitet aus dem Altgriechischem „διάγνωσις, *diágnosis*“ , bedeutet Unterscheidung, Entscheidung (bestehend aus διά-, *diá-*, ‚durch und γνώσις, *gnósis*, ‚Erkenntnis, Urteil‘) (Pape, 1914).

Hierdurch werden alle notwendigen Verfahren, die zur Erstellung einer Diagnose notwendig sind beschrieben. Die grundlegende Diagnostik besteht aus Anamnese und körperlicher Basisuntersuchung. Eine Anamnese zielt darauf ab, durch ein systematisches Gespräch mit dem Patienten möglichst viele Informationen zu den Beschwerden und dessen Ursachen gewinnen zu können. Reise- und Familienanamnese sowie Vorerkrankungen des Patienten spielen eine ebenso wichtige Rolle. Auf die Anamnese folgt eine körperliche Basisuntersuchung, die Vorgehensweisen wie Inspektion, Auskultation, Perkussion und Palpation beinhalten. Nach Berücksichtigung aller bekannten Einflussfaktoren auf die Beschwerden, wird eine Arbeits- bzw. Verdachtsdiagnose erstellt und diese durch die Gewinnung ergänzender Informationen, wie beispielsweise durch den Einsatz apparativer Methoden bestätigt, eingegrenzt oder verworfen. Zur gesicherten Diagnosestellung bedarf es häufig einer Wiederholung der oben genannten diagnostischen Schritte im zeitlichen Verlauf. Sollte durch die Wiederholungen die ursprünglich formulierte Hypothese nicht verifiziert werden können, wird der Diagnose ein „Verdacht auf“ vorangestellt. (Johner, 2018) Ebenso ist zu berücksichtigen, dass es andere Erkrankungen mit einem ähnlichen oder gleichen symptomatischen Bild gibt. Zieht man diese in Betracht, beschreibt man sie als sogenannte Differenzialdiagnosen. Die sogenannte Ausschlussdiagnose kommt über den schrittweisen Ausschluss verschiedener Erkrankungen zustande, bis letztendlich nur die Ausschlussdiagnose übrigbleibt. Kann keine Diagnose sicher festlegt werden, so gibt es die Möglichkeit den Patienten auf Verdacht zu therapieren. In Erwartung durch den Erfolg oder Misserfolg der Therapie eine Diagnose erstellen zu können, deklariert man diese Vorgehensweise als „Diagnosis ex juvantibus“. Hat man bereits eine Diagnose festgelegt und diese wurde wiederlegt, so beschreibt man diese als Fehldiagnose.

Unter Umständen sind die beiden oben genannten diagnostischen Verfahren (Anamnese und körperliche Untersuchung) nicht ausreichend, um einen gefährlichen Verlauf einer Erkrankung ausschließen zu können. Aus diesem Grund verwendet man daraufhin die apparative Diagnostik. Bei diesem Verfahren gilt es zu beachten, dass beim Einsatz der apparativen Diagnostik, der erwartete Nutzen dem Risiko der möglichen Beeinträchtigung des Patienten und den Kosten gegenübergestellt werden muss. Die Auswahl der diagnostischen Methoden sollte demnach unter Berücksichtigung der Wahrscheinlichkeit möglicher Therapieoptionen und potenzieller Gefahren der ausstehenden Differenzialdiagnosen entschieden werden. Grundsätzlich ist die apparative Diagnostik nur sinnvoll, wenn sich daraus Erkenntnisse für die Behandlung des Patienten gewinnen lassen. Die Diagnostik endet, wenn nur noch eine Diagnose in Frage kommt, oder eine weiterführende Diagnostik die Prognose eines ko- oder multimorbiden Patienten bzw. dessen Therapie nicht verändert. (Zetkin, 1980) Nach Festlegung einer Diagnose wird diese mittels ICD-10 Klassifikation beziffert und somit verschlüsselt. Das System der ICD-10 Klassifikation wird im folgenden Abschnitt kurz erläutert.

ICD-10 Klassifikation

Die Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitszustände (engl. International Statistical Classification of Diseases and Related Health Conditions, Abk. ICD) ist das weltweit anerkannteste und wichtigste Klassifikationssystem für medizinische Diagnosen. Die ICD-10 Klassifikation wurde durch das DIMDI (Deutsches Institut für Medizinische Dokumentation und Information) für Deutschland in eine deutsche Fassung namens ICD-10-GM überschrieben und dient der Verschlüsselung von Diagnosen bzw. Todesursachen. (Deutsches Institut für Medizinische Dokumentation und Information, 2018)

In Abbildung 1 wird exemplarisch das Kapitel der ICD-10-GM-2019 hinsichtlich der Diagnostik von Lungenrundherden dargestellt.

Kapitel XVIII
Symptome und abnorme klinische und Laborbefunde, die anderenorts nicht klassifiziert sind (R00-R99)

Abnorme Befunde ohne Vorliegen einer Diagnose bei bildgebender Diagnostik und Funktionsprüfungen
(R90-R94)

Inkl.: Unspezifische abnorme Befunde bei der bildgebenden Diagnostik:
- Computertomographie [CT]
- Kernspintomographie [MRI] [MRT] [NMR]
- Positronen-Emissions-Tomographie [PET]
- Röntgenuntersuchung
- Thermographie
- Ultraschall [Sonographie]

Exkl.: Abnorme Befunde bei der Screeninguntersuchung der Mutter zur pränatalen Diagnostik (O28.-)
Abnorme diagnostische Befunde, anderenorts klassifiziert - siehe Alphabetisches Verzeichnis

R90.- **Abnorme Befunde bei der bildgebenden Diagnostik des Zentralnervensystems**

R90.0 **Intrakranielle Raumforderung**

R90.8 **Sonstige abnorme Befunde bei der bildgebenden Diagnostik des Zentralnervensystems**
Abnormes Echoenzephalogramm
Krankheit der weißen Substanz o.n.A. [White matter disease]

R91 **Abnorme Befunde bei der bildgebenden Diagnostik der Lunge**
Inkl.: Lungenraumforderung o.n.A.
Rundherd o.n.A.

Abbildung 1: ICD-10-GM-2019 zum Stichwort Lungenrundherd in Anlehnung an (Deutsches Institut für Medizinische Dokumentation und Information, 2018)

2.2 Lungenrundherd

2.2.1 Definition

Ein Lungenrundherd ist eine rundliche bis ovale, intrapulmonal gelegene Struktur mit einem Durchmesser zwischen 1–3 cm. Ein Lungenrundherd größer als 3 cm wird als Raumforderung oder Tumor bezeichnet. (Tuddenham, 1984)
In einigen Definitionen werden hingegen auch Tumore mit bis zu 6 cm Durchmesser als Lungenrundherd bezeichnet, und auch das Kriterium „allseits von Lunge umgeben“ ist nicht in jeder Definition berücksichtigt (Hecker & Ukena, 2004).
In der vorliegenden Arbeit sollen unter dem Begriff Lungenrundherd keine Größenrestriktionen nach unten und oben hin berücksichtigt werden.
Per definitionem hat ein Lungenrundherd topographisch keinen Anschluss zum Hilus, Diaphragma, Brustwand oder Perikard. Dem hinzuzufügen ist, dass von dem als solitär definierten Lungenrundherd ausgehend keine Zeichen einer Begleitpneumonie, Atelektase oder regionären Lymphknotenvergrößerung vorliegen dürfen. (Tuddenham, 1984)

2.2.2 Radiologische Charakteristika & Symptomatik

Im Thoraxröntgenbild zeigt sich ein Lungenrundherd als rundliche Verschattung und muss in zwei Ebenen erkennbar sein. Liegt im Röntgenbild ein inhomogener und nicht

kreisrunder Lungenrundherd mit einer Größe von 1 bis 3 cm vor, so ist er als prinzipiell suspekt einzustufen. Dieser Verdacht gilt solange, bis das Gegenteil bewiesen ist. Lungenrundherde können solitär, multiple oder disseminiert vorliegen. Ein Lungenrundherd als solitär zu bezeichnen ist jedoch nur nach einer CT-Aufnahme sinnvoll, da durch dieses Verfahren über 50 % mehr Lungenrundherde, verglichen zum Thoraxröntgenbild, detektiert werden können. Zudem ist die Sensitivität, verglichen zum konventionellen Thoraxröntgenbild, bei der Detektion von Lungenrundherden < 1 cm in der CT deutlich höher, da hierbei eine genauere Abgrenzung zu den mediastinalen und hilären Strukturen gemacht werden kann. Lungenrundherde mit einem Durchmesser < 1 cm können im Thoraxröntgenbild meist nicht dargestellt werden. (Bergmann, Bölükbas, Beqiri, Trainer, & Schirren, 2007) Aufgrund der zweidimensionalen Abbildung eines Objekts im Thoraxröntgenbild können Überlagerungseffekte von Strukturen, die sich als Aufhellung oder Verschattungen zeigen, entstehen und die Präzision der Befundung limitieren. Daher ist häufig eine präzisere Diagnostik mittels CT-Aufnahme erforderlich. CT-Aufnahmen bieten die Möglichkeit aus zweidimensionalen Bildern dreidimensionale Datensätze zu errechnen und somit die Lokalisation von Strukturen in allen drei Ebenen darzustellen. (Bücheler, Lackner, & Thelen, 2005)

Die Erstellung eines konventionellen Röntgenbildes des Thorax weist eine Strahlenbelastung von 0,01 bis 0,1 Millisievert (Abk. mSv) auf. Die Strahlenbelastung einer CT-Aufnahme des Thoraxes beim Erwachsenen liegt hingegen zwischen 1 bis 10 mSv. (Heyer, 2007) Somit weist eine CT-Aufnahme des Thorax eine etwa 80-mal höhere Strahlenbelastung, verglichen zur konventionellen Röntgenbild-Aufnahme des Thorax in zwei Ebenen, auf (Bücheler et al., 2005).

Die kumulative Strahlenbelastung eines Menschen, bspw. durch natürliche Quellen, wie kosmische Strahlung, beträgt etwa 2,1 mSv pro Jahr. (Heyer, 2007) Folglich entspricht die Strahlenbelastung einer CT-Aufnahme ungefähr der Strahlenbelastung, welcher man auf natürlichem Wege über einen Zeitraum von vier Jahren ausgesetzt ist (Bücheler et al., 2005).

In der radiologischen Routine beruft man sich inzwischen am häufigsten auf die Kriterien der Fleischner Society um Lungenrundherde zu beurteilen und ggf. eine weitere Abklärung zu empfehlen. In Kapitel 2.2.4 werden die Empfehlungen der Fleischner Society zum Management von zufällig entdeckten, soliden und subsoliden Lungenrundherden bei Erwachsenen tabellarisch dargestellt.

Die Literatur „The solitary circumscribed pulmonary nodule“ nach GOOD, die auf bis zu 50-jähriger radiologischer Erfahrung basiert (Good & Wilson, 1958), bildet die Grundlage für die „2-Jahres-Regel“ nach HECKER (Hecker & Ukena, 2004). Diese sagt aus, dass ein Fehlen einer Größenzunahme eines Lungenrundherds im Thoraxröntgenbild als Hinweis für Benignität gilt, wenn diese über einen Zeitraum von mindestens 2 Jahren besteht (Good & Wilson, 1958). Demnach wird bei einem neu aufgetretenem Lungenrundherd im Röntgenbild empfohlen, die Voraufnahme der Patienten wiederholt zu

untersuchen. Sind in den letzten 2 Jahren keine radiologischen Größenzunahmen inspiziert worden, so sind keine weiteren diagnostischen Maßnahmen indiziert. Ist kein messbares Wachstum über 2 Jahre vorhanden, so spricht dieses Kriterium mit einem positiv-prädiktiven Wert von 65 % für Benignität. Vor dem Hintergrund, dass neuroendokrine Karzinome, als auch Bronchioloalveolarzellkarzinome über einen Zeitraum von mindestens 2 Jahren ein größenstabiles Verhalten aufzeigen, ist die zuvor gemachte Aussage jedoch als kritisch zu betrachten. Ebenso gilt es zu berücksichtigen, dass eine Größenänderung mindestens 3–5 mm betragen muss damit diese im Thoraxröntgenbild detektierbar ist. Durch das Auflösungsvermögen einer CT von 0,3 mm (HR-CT) ist dieses der konventionellen Röntgenuntersuchung deutlich überlegen. (Hecker & Ukena, 2004)

Abbildung 2 zeigt exemplarisch die vielfältigen radiologischen Morphologien, die ein Lungenrundherd aufweisen kann (Bergmann et al., 2007).

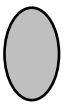
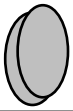
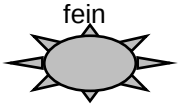
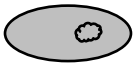
	Grundform	**Begrenzung**	**Verkalkung**	**Höhlenbildung/Kaverne**
rund		scharf 	zentral 	zentral 
oval		unscharf 	konzentrisch 	gekammert 
spindel		Corona radiata fein 	schollig (popcorn) 	exzentrisch 
asymme-trisch		Corona radiata grob 	diffuse 	dünnwandig 
gelappt		Pleurafinger 	exzentrisch 	+ Flüssigkeitsspiegel 

Abbildung 2: Beispiele für Lungenrundherde. Sie zeigen die Vielfältigkeit der radiologischen Morphologie des Lungenrundherdes in Anlehnung an (T. Bergmann, 2007)

Nach HECKER sind **Randbegrenzung bzw. Konfiguration des Lungenrundherdes** ein wichtiges diagnostisches Kriterium, um den Verdacht der Malignität eines Lungenrundherds einzugrenzen. Auch hier gilt das höhere Auflösungsvermögen der CT als vorteilhaft im Vergleich zur konventionellen Radiographie. Beispielsweise ist die *Corona radiata* (aus lat. Corona ‚Krone' und lat. radiatus ‚strahlend') als auch die

spikulierte (engl. spiculated) Läsion hinweisgebend für ein Karzinom. Eine spikulierte Läsion beschreibt eine Gewebeveränderung, die gezackte oder strahlige Ausläufer bildet. Ein gewellter Rand (engl. scalloped border) zeigen nach HECKER eine intermediäre Wahrscheinlichkeit für das Vorliegen eines Karzinoms. Ein glatter Rand hingegen spricht in der Regel für Benignität. (Hecker & Ukena, 2004)

Neben der Randbegrenzung bzw. Konfiguration des Lungenrundherdes ist nach HECKER auch die **Kalzifikation eines Lungenrundherdes** ein wichtiges Kriterium zum Ausschluss von Malignität (Hecker & Ukena, 2004). Liegt eine Kalzifikation vor, so spricht dies je nach anamnestischem Kontext oft eher für eine benigne Ätiologie. Auch hier gilt, dass die CT das Muster der Verkalkung besser als ein konventionelles Röntgenbild darstellt und eine Befundung des Lungenrundherds vereinfacht. (Berger, 2001)

Die verschiedenen **Muster der Verkalkung** sprechen für das Vorliegen von Lungenrundherden unterschiedlicher Genese. Beispielsweise ist eine laminierte oder zentrale Verkalkung ein typisches Merkmal für Granulome und eine weitere diagnostische Abklärung wird nicht empfohlen. Das klassische Popcornmuster hingegen ist häufig mit Hamartomen assoziiert. Bei malignen Erkrankungen sind exzentrische und getüpfelte (engl. stippled) Verkalkungsmuster ein Hinweis auf die Genese. (Tan, Flaherty, Kazerooni, & Iannettoni, 2003)

Aufgrund der unspezifischen **Symptomatik** bei Vorliegen eines Lungenrundherdes wird dieser meist als Zufallsbefund entdeckt. Das Vorliegen eines Lungenrundherdes kann asymptomatisch oder mit unspezifischen Symptomen, wie Brustschmerz, Hustenreiz oder Hämoptysen einhergehen. (Toomes, Delphendahl, Manke, & Vogt-Moykopf, 1983). Ca. 20–25 % aller Patienten geben keine charakteristischen Symptome an. Das Spektrum der zu berücksichtigenden Ursachen und Differenzialdiagnosen weist eine Zahl von bis zu 80 bekannten Erkrankungen auf, die im folgenden Unterkapitel beispielhaft genannt werden. Im Zusammenhang mit dieser Symptomatik ergibt sich die Diagnose eines Lungenrundherdes im Durchschnitt meist erst nach 7 Monaten. (Bergmann et al., 2007)

2.2.3 Genese und Differenzialdiagnose

Im folgenden Abschnitt wird die Genese und mögliche Differenzialdiagnosen von Lungenrundherden in Tabelle 1 dargestellt. Lungenrundherde können beispielsweise neoplastischer, infektiöser, inflammatorischer, vaskulärer, traumatischer oder kongenitaler Genese sein. Weitere benigne Differenzialdiagnosen können das Vorliegen von Rheumaknoten, intrapulmonale Lymphknoten, Plasmazellgranulome oder Sarkoidose sein. (Hecker & Ukena, 2004)

Tabelle 1: Differenzialdiagnose des Lungenrundherdes in Anlehnung an (Bergmann et al., 2007)

Entzündliche und narbige Rundherde	**Maligne Rundherde**	**Gefäßprozesse**
	Bronchialkarzinome	Arteriovenöse Fisteln
Tuberkulome	Metastasen	Varixknoten
Pneumonien	Sarkome	**Zysten**
Abszesse	Maligne Lymphome	
Kugelatelektasen	**Benigne und semimaligne Rundherde**	**Fremdkörper**
Septischer Embolus		
Eosinophile Infektion		Aspirierte
Aspergillome	Hamartochondrome	Äußerliche
Echinokokkus	Bronchusadenome	**Trauma**
Lues	Neurofibrome, Fibrome	
	Lipome	Kontusionsblutung
	Osteome	Blutung nach Biopsie (iatrogen)

Das ausgeprägt Blut- und Lymphgefäßsystem zum Hals, als auch zum Bauch lassen die Lunge ein häufiges Zielorgan für Metastasen sein. 50 % der obduzierten Patienten, die an einem Tumor verstorben sind, zeigten Metastasen in der Lunge. Die häufigsten Primärtumore finden sich in Geweben der Mamma, Niere, Kopf, Hals, sowie Gastrointestinaltrakt und der Haut. (Adler, 2011)

2.2.4 Diagnostik

Im folgenden Abschnitt werden die diagnostischen Schritte zur Detektion und Evaluierung von Lungenrundherden beschrieben. Exemplarisch zeigt Abbildung 3 einen Überblick zur Abklärung pulmonaler Rundherde. Weitere orientierende Veröffentlichungen mit Richtlinien und Vorschlägen zur Abklärung und Therapie von Lungenrundherden sind die Fleischner Society, sowie die AWMF-Leitlinie. (Adler, 2011)

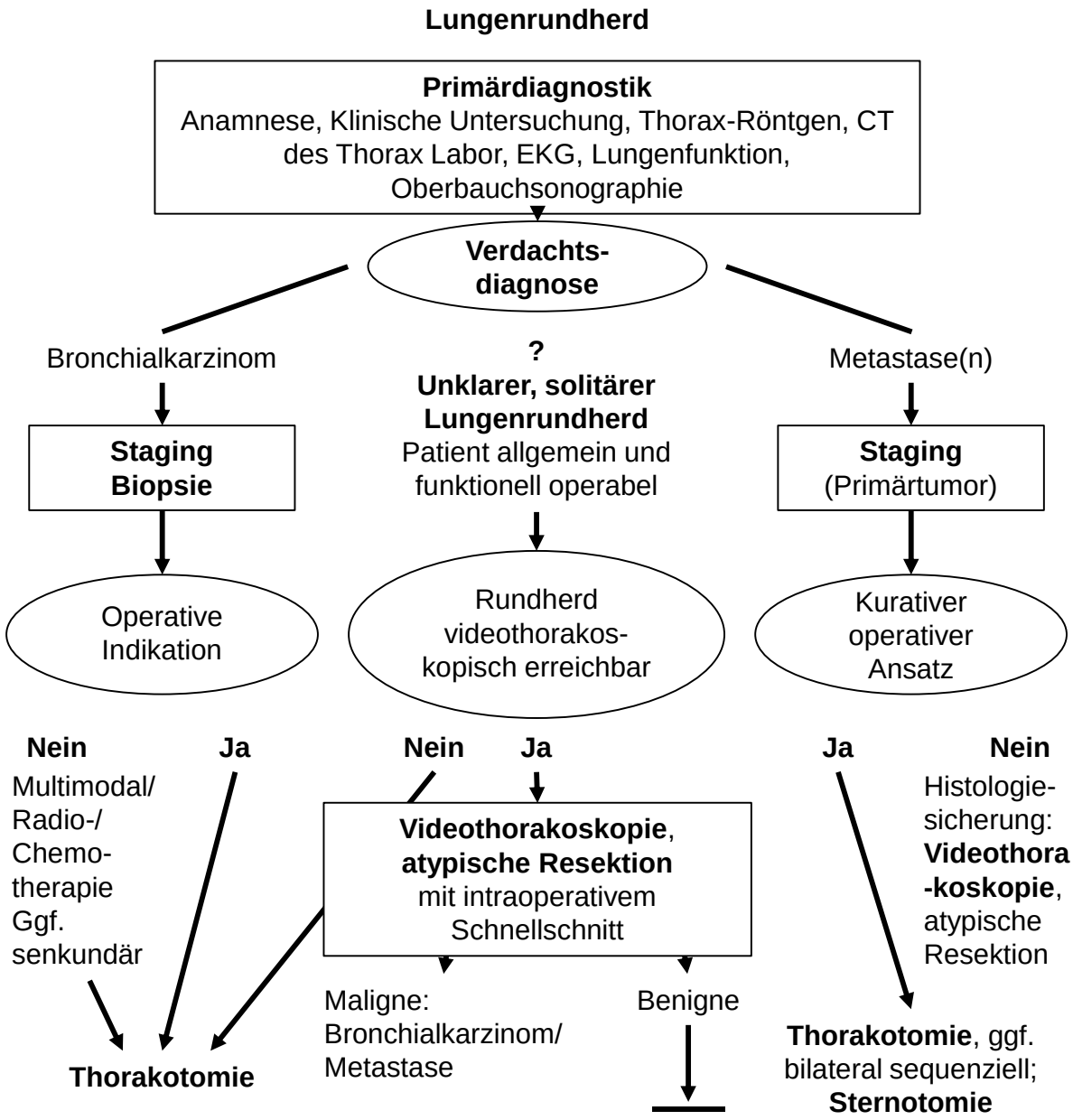


Abbildung 3: Algorithmus zur Abklärung pulmonaler Rundherde in Anlehnung an (Hoffmann & Dienemann, 2000)

In den Empfehlungen der Fleischner Society (dt. Fleischner-Gesellschaft) von 2017 zeigt sich, wie in Tabelle 2 dargestellt, ein individuelles weiteres diagnostisches Vorgehen in Abhängigkeit von der Rundherdgröße, dem singulär oder multiplen Vorliegen dieser und dem individuellen Risiko des Patienten (MacMahon et al., 2017). Die Fleischner Society ist eine internationale, multidisziplinäre medizinische Gesellschaft für Thoraxradiologie, die Empfehlungen bzw. Leitlinien zur Diagnose und Behandlung von verschiedenen Erkrankungen des Brustkorbs erstellt.

Tabelle 2: Leitlinien der Fleischner Society 2017 zum Management von zufällig entdeckten, soliden und subsoliden Lungenrundherden bei Erwachsenen in Anlehnung an (MacMahon et al., 2017)

Rundherdgröße	Vorliegen solider Lungenrundherde	CT-Kontrollempfehlung bei Patient mit niedrigem Risiko	CT-Kontrollempfehlung bei Patient mit hohem Risiko
< 6 mm	singulär	Keine	Optional in 12 Monaten
	multiple	Keine	Optional in 12 Monaten
6–8 mm	singulär	in 6–12 Monaten, danach CT in 18–24 Monaten erwägen	in 6–12 Monaten, danach CT in 18–24 Monaten
	multiple	in 3–6 Monaten, danach CT in 18–24 Monaten erwägen	in 3–6 Monaten, danach CT in 18–24 Monaten
> 8 mm	singulär	Kontrolle in 3 Monaten erwägen, weitere Diagnostik mittels PET/CT oder Biopsie	Kontrolle in 3 Monaten erwägen, weitere Diagnostik mittels PET/CT oder Biopsie
	multiple	Kontrolle in 3–6 Monaten, Kontrolle nach 18–24 Monaten erwägen	Kontrolle in 3–6 Monaten, Kontrolle nach 18–24 Monaten

Rundherdgröße	Vorliegen subsolider Lungenrundherde	CT-Kontrollempfehlung
< 6 mm	Milchglasherd, solide	Keine
	teil-solide	Keine
	multiple	Kontrolle nach 3–6 Monaten, Kontrolle nach 2 und 4 Jahren erwägen, falls stabiler Befund
≥ 6 mm	Milchglasherd, solide	Kontrolle nach 6–12 Monaten, bei Persistenz Kontrolle alle 2 Jahre über 5 Jahre
	teil-solide	Kontrolle nach 3–6 Monaten, bei Persistenz jährliche Kontrolle für 5 Jahre
	multiple	Kontrolle nach 3–6 Monaten, anschließendes Management basierend auf verdächtigstem Herd

Zur Diagnostik gehören, wie bereits allgemein in Kapitel 2.1 erläutert, eine ausführliche Anamnese, bei der insbesondere das Geschlecht, Alter, Gewichtsverlauf, maligne Vor- und Begleiterkrankung, Familienanamnese, berufliche Gefahrstoffexposition und Tabakkonsum des Patienten erhoben werden. Es folgen klinische Untersuchungen und bei Verdacht auf ein pulmonales Geschehen die Anfertigung von Thoraxröntgenbildern in 2 Ebenen, Computertomographie des Thorax, sowie die Lungenfunktionsdiagnostik. (Bergmann et al., 2007)

Bei bereits bestehenden Lungenrundherden, die in der Vergangenheit als benigne klassifiziert wurden, können Veränderungen in Größe und Form können Wegweiser für ein malignes Geschehen sein und sind weiter zu untersuchen (Hecker & Ukena, 2004). Abbildung 4 zeigt zudem eine enge Korrelation zwischen Größe des Lungenrundherdes und dessen Bösartigkeit.

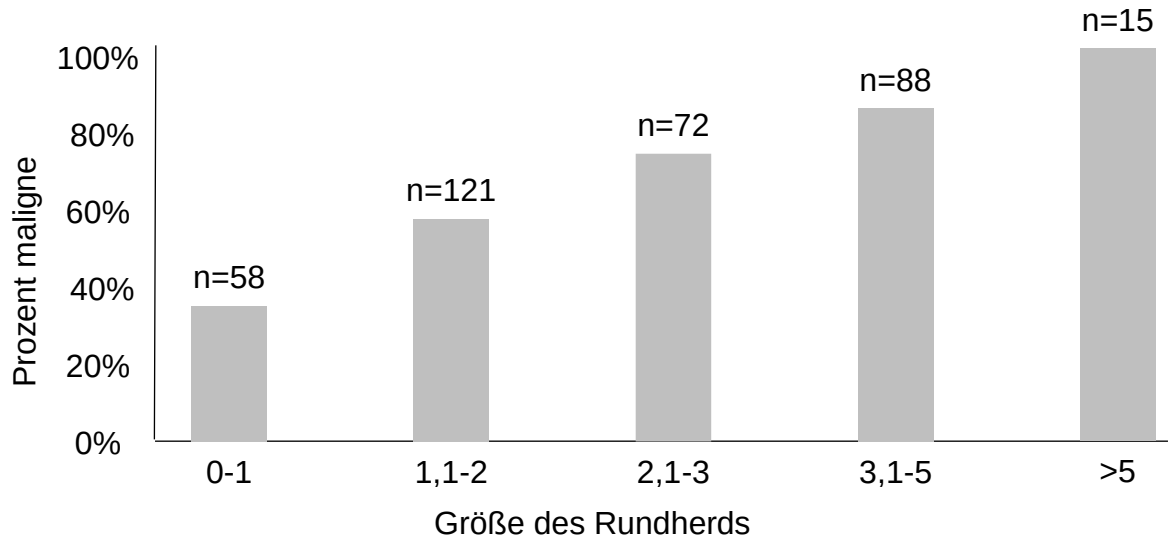


Abbildung 4: Inzidenz für Bösartigkeit eines solitären Rundherdes in Abhängigkeit von der Rundherd-Größe (n = 360) in Anlehnung an (Bergmann et al., 2007)

Nach LILLINGTON bedeutet ein Zuwachs von 28 % im Durchmesser des Lungenrundherds gleichzeitig eine Verdopplung der Tumormasse. Untersuchungen von malignen Geschehen haben gezeigt, dass dieses Wachstum in einem Zeitraum von 25–450 Tagen geschieht. Der Median liegt bei 120 Tagen. (Lillington, 1991)

Zu unterscheiden ist die Tumorverdopplungsrate von kleinzelligen zu nicht-kleinzelligen Lungenkarzinomen, die individuell eine große Variationsbreite aufweist. Das kleinzellige Lungenkarzinom hat eine Tumorverdopplungsrate von ca. 30 Tagen, wohingegen diese bei nicht-kleinzelligen Lungenkarzinomen bei 100 Tagen liegt. (Geddes, 1979)

Die Abschätzung des Lungenrundherd-Durchmessers ist mit einer 10%igen Variationsbreite verbunden. Ein langsames Wachstum stellt eine besondere Schwierigkeit in der Detektion im Standard-CT dar. Dies berücksichtigend lässt sich nachvollziehen, warum die Beurteilung über das Wachstum bzw. die Stabilität eines Herdes sehr

schwierig ist und eine lange Beobachtungszeit benötigt. (Brenner, Holsti, & Perttala, 1967)

Zur Sicherung der Verdachtsdiagnose und Feststellung der Genese des Lungenrundherdes ist ein invasives Verfahren notwendig. Die Benignität oder Malignität kann mit keinem nicht-invasiven diagnostischen Verfahren zweifelsfrei dargestellt werden. Aus diesem Grund dient der Einsatz der unterschiedlichen bildgebenden Verfahren zur Lokalisation, Darstellung und Operationsplanung des Herdes. Bei Verdacht auf eine maligne tumoröse Raumforderung in der Lunge wird diese in der Regel histologisch gesichert. Hierbei kann es zudem sinnvoll sein, zum Beispiel leicht zugängliche, auffällige, pathologisch vergrößerte Lymphknoten zu biopsieren und anschließend ebenfalls histologisch zu untersuchen. Diese Methode ist die für den Patienten der kleinstmögliche invasive Eingriff, der zur Sicherung oder Eingrenzung eines malignen Prozesses dient. Sind keine auffälligen Lymphknoten vorhanden oder ist die histologische Untersuchung negativ ausgefallen, so ist der nächste mögliche diagnostische Schritt die direkte Biopsie des Lungenrundherds. In jeder Hinsicht müssen der Patientenwunsch und seine persönliche aktive Entscheidung in das weitere Prozedere einbezogen werden. Eine umgehende Diagnosestellung hat signifikanten Einfluss auf die Heilungschancen, denn das Tumorstadium korreliert eindeutig mit der Prognose eines Bronchialkarzinoms. Aus onkologischer Sicht geht die chirurgische Resektion eines isolierten Lungenrundherdes (Tumorstadium 1a) mit einer 80%igen 5-Jahres Überlebensrate einher. Die definitive Diagnose ist jedoch erst nach pathologischer Untersuchung als gesichert anzunehmen. (Ost, Fein, & Feinsilver, 2003; Tan et al., 2003)

Nach HOFFMANN & DIENEMANN ist die 5-Jahres-Überlebensrate umso größer, je kleiner der Tumor und je günstiger der Lymphknotenstatus zum Zeitpunkt der Operation ist (Hoffmann & Dienemann, 2000).

Tabelle 3 zeigt die Zusammenhänge zwischen Tumorgröße und Heilungsrate.

Tabelle 3: Zusammenhänge zwischen Tumorgröße und Heilungsrate in Anlehnung an (Hecker & Ukena, 2004)

Diameter in mm	Anteil in %	Heilungsrate[a] in %
5–15	10	69 (64–74)
16–25	30	63 (60–67)
26–35	26	58 (54–61)
36–45	15	53 (48–57)
> 45	19	43 (39–48)

[a] *12 Jahres-Überlebensrate*

Die unterschiedlichen Biopsieverfahren weisen eine unterschiedliche Sensitivität zur Identifikation maligner und benigner Herde auf. Sie sind abhängig von technischen Voraussetzungen, wie Kaliberstärke und Biopsienadeln und der Menge und Qualität des entnommenen Materials, das zur pathologischen Befundung dient. Nach HOFFMANN kann der Anteil unzureichenden Materials einer Biopsie bis zu 28 % betragen. (Hoffmann & Dienemann, 2000) Aus diesem Grund lässt sich die Abwesenheit maligner Zellen nicht eindeutig als Beweis eines benignen Geschehens heranziehen. (Gasparini et al., 1995; Lillington, 1991)

Bei der transbronchialen Biopsie (TBB) ist die Sensitivität der Biopsie bei kleinen peripheren pulmonalen Rundherden gering und liegt bei unter 30 %. Wird dieses Verfahren jedoch mit der CT-gesteuerten transthorakalen-perkutanen Nadelbiopsie (TTNB) kombiniert, so kann die Sensitivität zur Identifikation maligner Herde auf etwa 90-99 % gesteigert werden. (Gasparini et al., 1995) Für benigne Herde ist jedoch die diagnostische Sicherheit zur Identifikation solcher Herde wesentlich geringer und wird zwischen 5-89 % angegeben (Peuchot & Libshitz, 1987).

Abbildung 5 zeigt die prozentuale Verteilung von 955 operierten Lungenrundherden und ihrer Dignität im Zeitraum von 1970 bis 1980 in Mitteleuropa, basierend auf der Datenerhebung von Toomes et al.

Diese Untersuchung zeigt, dass ca. 50 % der Lungenrundherde als maligne einzustufen waren. Gleichzeitig beschrieb Toomes et al. dass diese Daten abhängig von dem untersuchten Patientenkollektiv, den geographisch bedingten Unterschieden, als auch dem klinischen Material sind. (Toomes et al., 1983)

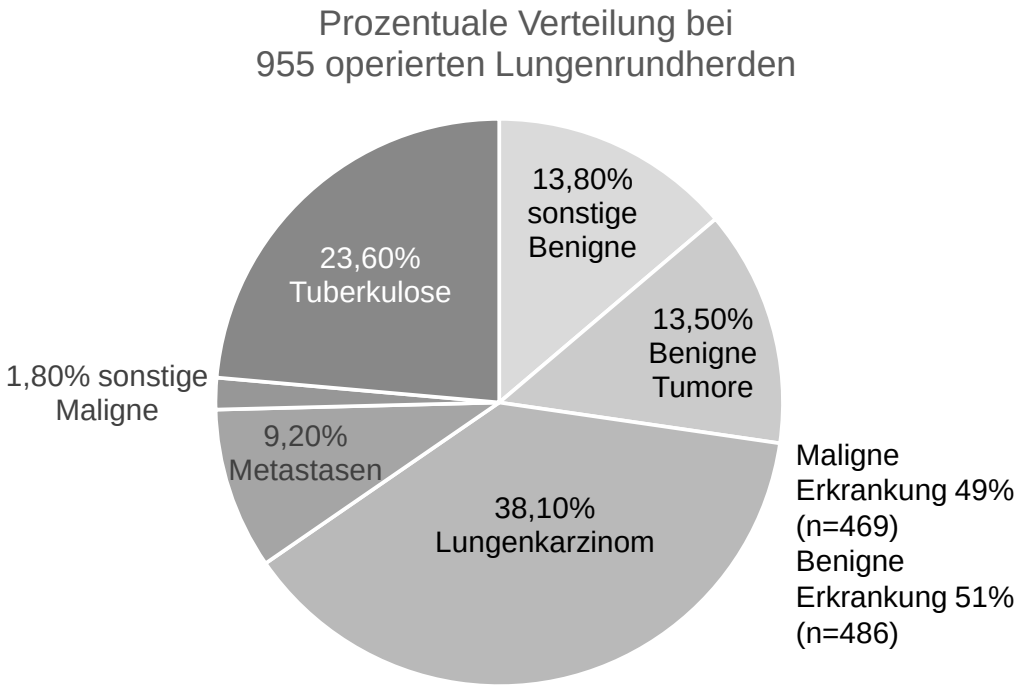


Abbildung 5: Prozentuale Verteilung bei 955 operierten Lungenrundherden im Zeitraum von 1970 bis 1980 in Mitteleuropa in Anlehnung an (Bergmann et al., 2007)

2.2.5 Risikostratifizierung und Inzidenz

Die Wahrscheinlichkeit, dass es sich bei einem im Röntgenbild inhomogenen und nicht kreisrunden Lungenrundherd mit dem Durchmesser von 1-3 cm, um einen malignen Herd handelt, nimmt mit Zunahme der Risikofaktoren und Ausgangsgröße des Herdes zu (vgl.

Tabelle 4). Personen aus Regionen mit einer überdurchschnittlichen Zahl an pilz- und parasitären Infektionen oder Patienten mit einer aktuell malignen Erkrankung bzw. malignen Vorgeschichte sind besonders für die Entstehung von Lungenrundherden prädestiniert. Ebenso stellen Personen höheren Alters, sowie Raucher eine Gruppe mit erhöhtem Risiko dar. (Bergmann et al., 2007)

Tabelle 4 und Tabelle 5 geben einen Überblick über die Indizien und eine Risikostratifizierung zur Dignitätsbeurteilung von Lungenrundherden.

Tabelle 4: Indizien zur Dignitätsbeurteilung pulmonaler Rundherde (RH) in Anlehnung an (Hoffmann & Dienemann, 2000)

Wahrscheinlich benigne	**Wahrscheinlich maligne**
Junger Patient (< 35 Jahre)	Älterer Patient (> 35 Jahre)
Nichtraucher	Raucher
Keine Malignomanamnese	Malignomanamnese
RH-Durchmesser < 3 cm	RH > 3 cm
RH verkalkt	RH nicht verkalkt
RH glatt begrenzt	RH unscharf begrenzt
Keine Größenzunahme d. RH in 2 Jahren	Nachgewiesene Größenzunahme

Tabelle 5: Risikostratifizierung von Patienten mit isolierten Lungenrundherden in Anlehnung an (Hecker & Ukena, 2004)

Variable	Malignomrisiko		
	Niedrig	**Mittel**	**Hoch**
SPN-Durchmesser	< 1,5 cm	1,5–2,2 cm	> 2,3 cm
Alter	< 45 Jahre	45–60 Jahre	> 60 Jahre
Raucherstatus	Nichtraucher	Aktueller Raucher[a]	Aktueller Raucher[b]
Stopp Nikotinkonsum	Vor > 7 Jahren[c]	< 7 Jahren	Niemals
SPN-Charakteristika	Glatt	gewellt	*Corona radiata* od. spikuliert

[a] < 20 Zigaretten/Tag; [b] > 20 Zigaretten/Tag; [c] oder Nichtraucher

Nach HECKER liegt zu etwa in einem Drittel aller Fälle ein primäres Malignom in der Erscheinung eines solchen solitären Lungenrundherds vor. In ca. 25 % aller Fälle handelt sich um eine solitäre Metastase. (Hecker & Ukena, 2004)
BERGMANN gibt an, dass die Wahrscheinlichkeit, dass es sich bei einem 65 Jahre alten, langjährigen Raucher, um einen malignen Herd handelt, bis zu 95 % betragen kann. Zudem sagt BERGMANN aus, dass die Prävalenz und auch die Inzidenz des Lungenrundherdes in der Gesamtbevölkerung nicht genau bekannt ist, da dies zum einen abhängig von der Definition, als auch von der Personengruppe und der Sensitivität der Diagnostik ist. (Bergmann et al., 2007)
Weiterführend kann nach HECKER durchschnittlich ein solitärer Lungenrundherd pro 500 Röntgenbildern des Thorax beobachtet werden. 90 % dieser Lungenrundherde sind Zufallsbefunde. (Hecker & Ukena, 2004) LILLINGTON's Untersuchungen ergaben, dass in den USA jährlich bei ca. 150.000 Patienten, ein Lungenrundherd neu diagnostiziert wird. 40–50 % aller neu diagnostizierten Lungenrundherde gelten als maligne. (Lillington, 1990)
All diese prozentualen Verteilungen sind jedoch abhängig von der berücksichtigten Personengruppe, sowie der Sensitivität der bildgebenden Diagnostik. Bei einem Patienten mit positiver Raucheranamnese, älter als 65 Jahren und ggf. dem Vorliegen von Vorerkrankungen muss man mit einer Wahrscheinlichkeit von 95 % mit einem malignen Prozess rechnen. Ein Patient im Alter von 35 Jahren ohne Raucheranamnese hat ein Risiko für einen malignen, peripheren, solitären Rundherd von ca. 1 % (vgl. Abbildung 6 nach Bergmann). Die Wahrscheinlichkeit für Malignität steigt mit Zunahme des

Alters (vgl. Abbildung 6) und wachsender Anzahl in Frage kommender Risikofaktoren rapide an. (Rubins & Bloomfield Rubins, 1996)

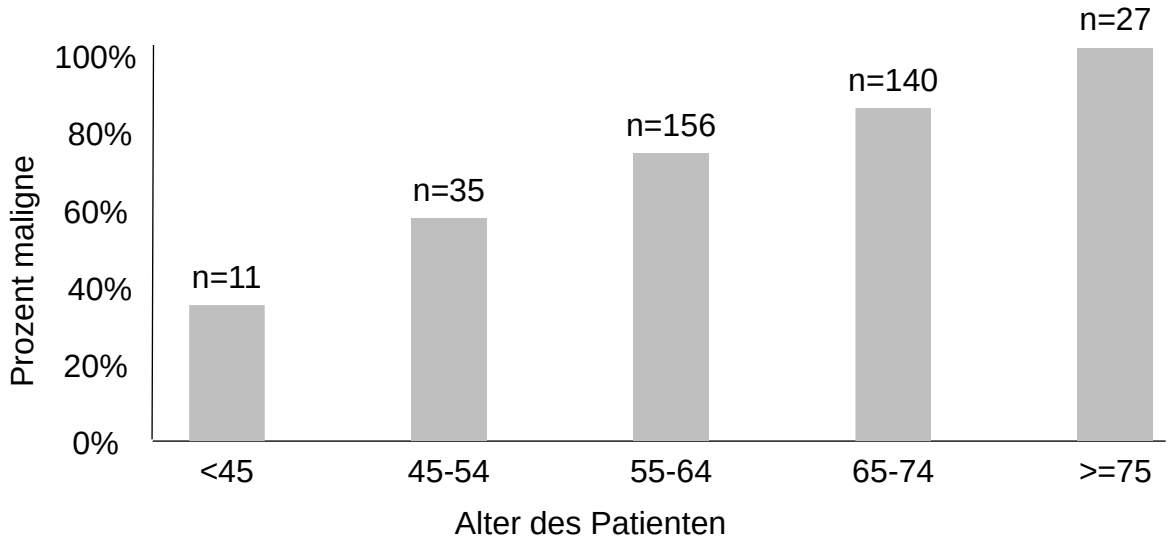


Abbildung 6: Inzidenz für Bösartigkeit eines solitären Rundherds in Abhängigkeit vom Alter (n = 360) in Anlehnung an (Bergmann et al., 2007)

2.3 Künstliche Intelligenz

Im Folgenden werden die Definition und die Anwendungsbereiche der künstlichen Intelligenz, sowie des maschinellen Lernens beschrieben.

2.3.1 Definition

Künstliche Intelligenz (Abk. KI; engl.: artificial intelligence*,* Abk. AI) ist ein Teilgebiet der Informatik. Der Begriff wurde erstmals 1956 von dem amerikanischen Informatiker John McCarthy (*1927) benutzt und umfasst zwei wesentliche Aspekte: die Erschaffung und Nachahmung menschlichen Verhaltens und Denkens sowie die Integration automatischer und autonomer Aufgabenbewältigung.

Die angestrebten Ziele der KI sind unter anderem die Mechanisierung menschlichen Denkens oder die Nachahmung menschlichen Verhaltens, wobei letzteres als sehr anspruchsvoll und derzeit noch visionär einzuschätzen ist. Die Übernahme von eindeutig und klar definierten Aufgabenbereichen durch eine intelligente Maschine hingegen findet derzeit schon Einsatz in verschiedenen Gebieten statt, die im nächsten Kapitel genauer erläutert werden.

Derzeit ist der Begriff der künstlichen Intelligenz noch nicht ausreichend definiert, dennoch wird dieser in der Forschung und Entwicklung bereits verwendet. (Goram, 2018a; Wichert, 2018) In diesem Zusammenhang muss der Begriff kritisch betrachtet werden. Die nachfolgende Definition der KI nach CASTRO & NEW soll als Grundlage für die vorliegende Arbeit benutzt werden: KI ist ein Bereich der Informatik, der sich mit der Entwicklung von Maschinen und Systemen beschäftigt, die Operationen analog zum

menschlichen Lernen, Handeln und Entscheiden durchführen. Hierbei sind die wesentlichen Funktionalitäten von KI-Systemen als Lernen, Verstehen, Schlussfolgern, Urteilen und gelungener Interaktion zwischen Maschine und Mensch definiert. (Castro & New, 2016)
Die KI verwendet die symbolische Wissensrepräsentation zur Konzeption und Nutzung von Systemen. Hierbei speichert das Programm Wissen, Fakten und Regeln zu einem bestimmten Thema. Diese werden anschließend mit sogenannten „*Wenn-dann-Regeln* (*wenn* eine bestimmte Bedingung auftritt, *dann* handle auf eine bestimmte Art und Weise)“ spezifiziert. (Hanser & Scholtyssek, 2000)
Die KI ist ein Bereich, der durch verschiedene Fachbereiche entstanden ist und durch den technischen Aspekt des Fachbereichs der Informatik verknüpft wurde. Hierbei handelt es sich beispielsweise um Themenfelder, wie Neurowissenschaften, Psychologie, Mathematik, Philosophie, Linguistik und auch Kommunikationswissenschaften. (Goram, 2018a)

2.3.2 Anwendungsbereiche

Im Folgenden werden einige Anwendungsbeispiele der KI angeführt, um die Vielfältigkeit dieses Themengebiets zu verdeutlichen.
Das teilautonome Fahren, intelligente Softwareassistenten, automatische Sprach- und Bilderkennung sind Systeme, die aus dem Alltag nicht mehr wegzudenken sind. Die Mustererkennung ermöglicht eine automatische Auswertung und Klassifizierung von Daten. Dies wird beispielsweise bei der Analyse des Kaufverhaltens von Kunden in Supermärkten eingesetzt. Auch die Gesichtserkennung spielt in der Forensik mittlerweile eine entscheidende Rolle. Automatisierte Spracherkennung, welche Sprache ohne Zeitverzögerung übersetzen können soll, befindet sich derzeit noch in der Entwicklung. Des Weiteren wird KI in komplexe Optimierungsprobleme wie zum Beispiel Logistikplanungen eingebunden, um bei dessen Bewältigung zu unterstützen. Die Zuverlässigkeit dieser Systeme wird durch die fortlaufende Sammlung neuer Umwelteinflüsse verbessert. (Goram, 2018a)
Auf die Anwendungsbeispiele von künstlicher Intelligenz und maschinellem Lernen im medizinischen Bereich wird in Kapitel 2.4.2 näher eingegangen.

2.3.3 Turing-Test

Der britische Mathematiker A. M. Turing (1912–1954) schrieb 1950 einen für die künstliche Intelligenz entscheidenden Aufsatz, der den Titel „Computing Machinery and Intelligence" trägt (Turing, 2021). Zur Feststellung, inwiefern das Verhalten der KI mit dem Abzubildenden übereinstimmt, wurde der *Turing-Test* entwickelt. Bei diesem Test wird eine Testperson einem Programm und einem Menschen verdeckt gegenübergestellt. Das Programm wird als intelligent bezeichnet, sollte die Testperson bei der Kommunikation nicht zwischen Mensch und Programm unterscheiden können. In der Regel

sind die Gesprächsthemen bei dem Turing-Test uneingeschränkt wählbar. Belaufen sich die gestellten Fragen thematisch nur auf ein Gebiet, so wird von einem eingeschränkten Turing-Test gesprochen (bspw. Schachspielen oder medizinische Diagnose). Die Wahrnehmung wird bei diesem Test nicht berücksichtigt, sondern vielmehr die symbolische Wissensrepräsentation. (Wichert, 2018)

2.3.4 Schwache künstliche Intelligenz

Die *schwache KI* (engl. weak bzw. narrow AI) basiert auf der Simulation von Intelligenz und datenbankgesteuerten, erlernten Reaktionen (Vehmeyer, 2014). Die schwache KI hat als Ziel, Lösungen für klar definierte Problemstellungen mittels einer konkreten Herangehensweise zu finden. Hierbei kann die schwache KI nicht auf selbstständig erarbeitete Methoden zurückgreifen, sondern verwendet ausschließlich die ihr zu Verfügung gestellten Ressourcen. Das System der schwachen KI arbeitet reaktiv innerhalb eines begrenzten Intelligenz-Levels und hat kein weitreichendes Verständnis für die Problemlösung. Dennoch ist das System in der Lage sich selbst zu optimieren. Die derzeit existierenden Systeme, welche bereits im Alltag Anwendung finden, sind nahezu ausschließlich der schwachen KI zuzuordnen. Beispiele sind Navigationssysteme, Text-, Bild-, Zeichen-, Spracherkennung, automatisierte Übersetzung, individuelle Werbung, Autovervollständigung, oder Korrekturvorschläge bei Suchvorgängen und Tippfehlern. Eine rasche Entwicklung zeigt sich besonders in Verwaltungsaufgaben, wie Kundensupport mit Chatforen oder BackOffice Aufgaben. (Moeser, 2017)

2.3.5 Starke künstliche Intelligenz

Die *starke KI* (engl. strong bzw. general AI) wird auch „Superintelligenz“ genannt (Moeser, 2017). Unter diesem Begriff werden jene Ansätze zusammengefasst, die einer Abbildung des menschlichen Gehirns und den damit verbundenen menschlichen Vorgängen und Handlungsweisen am nächsten kommen, diese zu imitieren (Buxmann & Schmidt, 2019), oder sogar zu übertreffen versuchen (Moeser, 2017).
Nach MOESER ist die Mehrheit der Forscher der Meinung, dass die Entwicklung eines mit menschlichen, kognitiven Fähigkeiten ausgestatteten Systems in den kommenden 20–40 Jahren generell möglich sei. Forscher sind sich jedoch bereits jetzt schon einig, welche Kriterien die starke KI erfüllen muss, um als solche bezeichnet werden zu können. Hierbei handelt es sich um die Fähigkeit, das logische Denkvermögen einzusetzen und Entscheidungen treffen zu können, auch bei dem Vorliegen von Unsicherheit. Ein weiterer wichtiger Aspekt ist die Eigenschaft der Planungs- und Lernfähigkeit und die Voraussetzung all dies in natürlicher Sprache zu kommunizieren. In diesem Zusammenhang soll das Erreichen eines übergeordneten Ziels durch die Kombination aller Fähigkeiten gegeben sein. (Moeser, 2017) Zudem erwartet man von der starken KI, dass diese die Fähigkeit aufweist, sich autonom weiterzuentwickeln und selbstständig zu lernen (Vehmeyer, 2014).

Nach MOESER steht eine starke KI in besonderem Zusammenhang mit nicht mehr nur reaktiven, sondern vielmehr intelligenten und flexiblen Handlungen (Moeser, 2017).
Als konstituierende Merkmale umfasst diese die Eigenschaften, wie Bewusstsein und Empathie (Goertzel & Pennachin, 2007; Searle, 1980). Diese Eigenschaften sind jedoch noch Gegenstand der heutigen Forschung (Buxmann & Schmidt, 2019).
Ob die KI überhaupt in der Lage ist die menschlichen Fähigkeiten, wie Gedächtnis, Bewusstsein, Weisheit und die echte Empathie zu erlangen, oder ob eine zweckgebundene simulierte Empathie erlernt werden kann, wird derzeit noch spekuliert (Moeser, 2017).

2.4 Maschinelles Lernen

2.4.1 Definition

Maschinelles Lernen (engl. Machine Learning) ist ein Teilgebiet der KI und beschreibt als Oberbegriff das Erlangen von Wissen und Fähigkeiten aus bestehenden Erfahrungen heraus. Es basiert auf der Methodik, das computergestützte Systeme anhand spezifischer Algorithmen und Beispiele trainiert werden. (Goram, 2018a, 2018b) Die, für die erstmalige Funktionserstellung notwendigen, musterhaften Daten, müssen anhand einheitlicher Regeln vorab händisch auserwählt werden. Anschließend werden die vorhandenen Daten in drei verschieden Gruppen aufgeteilt, die zu verschiedenen Zeitpunkten des Lernprozesses verwendet werden. Zu Beginn arbeitet das System mit Trainingsdaten die dazu dienen eine Funktion mit den entsprechenden Parametern zu definieren. Anschließend wird ein Datensatz zur Validierung hinzugezogen, um die Richtigkeit der Funktion zu prüfen. Abschließend prüfen die Testdaten die Performance und Genauigkeit der Funktion. (Scherk, Pöchhacker-Tröscher, & Wagner, 2017)
Nach Durchlaufen dieses Lernprozesses ist das System in der Lage Muster und Gesetzmäßigkeiten zu erkennen und diese auf unbekannte Daten anzuwenden. Dieser Vorgang wird als Lerntransfer bezeichnet. (Bishop, 2006)
Das Ziel von maschinellem Lernen ist das eigenständige Lösen von Problemen. Derzeit ist maschinelles Lernen daher ein interdisziplinäres Forschungsfeld, das „Sprach- und Computerwissenschaften, sowie Lernpsychologie vereint“. (Goram, 2018b)

2.4.2 Anwendungsbereiche

Nach LITZEL kann maschinelles Lernen mittlerweile ein breites Spektrum an Aufgaben erfüllen. Die Systeme des maschinellen Lernens sind in der Lage, für ihre Aufgaben relevante Daten zu finden, zu exportieren und anschließend auf verallgemeinerte Anwendungsgebiete anzuwenden. Darüber hinaus ist es dem System möglich, basierend auf den analysierten Daten, Vorhersagen und Wahrscheinlichkeiten für bestimmte Ereignisse zu bestimmen. Des Weiteren können sich die auf maschinellem Lernen

basierenden Systeme selbstständig an Entwicklungen anpassen und durch die Erkennung von Mustern sich stetig weiterentwickeln. (Litzel, 2016)
Derzeit wird maschinelles Lernen bei der Erkennung von Spam-Emails, als Spamfilter in Email-Accounts (Litzel, 2016) oder in der Analyse des Such- und Kaufverhaltens mit anschließend individualisierten Produktvorschlägen verwendet (Goram, 2018b). Es findet auch Anwendung bei digitalen Assistenten in der Sprach-, Bild-, Gesichts- und Texterkennung, sowie der Entscheidung von Relevanz von Webseiten unter der Berücksichtigung bestimmter Suchbegriffe. Maschinelles Lernen ist ebenfalls in der Lage anhand des Verhaltens, die Internetaktivität von natürlichen Personen und Bots (von engl. robot, dt. Roboter) zu differenzieren. In der Forensik spielt es eine wichtige Rolle in der automatischen Erkennung von Kreditkartenbetrug. (Litzel, 2016)
Maschinelles Lernen wird in der Medizin unter anderem zur Analyse der Bedeutung klinischer Parameter und ihrer Kombinationen für die Erstellung einer Prognose, als auch die Vorhersage des Krankheitsverlaufs, Gewinnung von medizinischem Wissen für die Ergebnisforschung, Therapieplanung und -unterstützung, sowie für das verbundene Patienten-Management eingesetzt. Ein weiterer Aspekt der Verwendung von maschinellem Lernen ist die Datenanalyse. Hierbei können Regeln erkannt werden, selbst wenn Daten unvollständig sind. Ebenso kann die Interpretation der kontinuierlichen Daten, die auf Intensivstationen gesammelt werden, mit einer intelligenten Alarmierung zu einer effektiven und effizienten Überwachung des Patienten dienen. (Magoulas & Prentza, 2001) Zudem befasst es sich mit der Identifizierung von Krankheiten und Diagnosestellung, als auch geeigneter Wirkstofffindung und -herstellung. Aktuell liegt der Fokus des maschinellen Lernens sowohl auf der personalisierten Medizin, Analyse und Diagnosestellung in der medizinischen Bildgebung, als auch auf Verhaltensanalysen zur Prävention von Krankheit und Förderung von Gesundheit. Zudem sind Krankenakten, klinische Studien und Forschungen, als auch crowdsourced Datenerfassung und verbesserte Strahlentherapie, als auch Prognose von Krankheitsausbrüchen von großer Bedeutung. (Flatworldsolutions, 2020) Konkretere Anwendungsfälle des maschinellen Lernens sind beispielsweise die Erkennung von Lungenkrebs oder Schlaganfällen auf Basis von CT-Aufnahmen, sowie die Beurteilung des Risikos eines plötzlichen Herztodes oder anderer Herzerkrankungen anhand von Elektrokardiogrammen und Herz-MRT-Aufnahmen. Ebenso können Hautläsionen in Hautbildern klassifiziert und Indikatoren für diabetische Retinopathie in Augenbildern gefunden werden. (Schmitt, 2019) Auch bei dem Einsatz der Erkennung von Lungenkrebs anhand histopathologischer Diagnosen, zeigte sich, dass KI-Algorithmen sich als wirksamer erwiesen haben als ein Mensch. (Yu et al., 2016) Aber auch in der Arzneimittelentwicklung kann es erfolgreich eingesetzt werden. Hierbei wird es bei der Identifizierung von Interventionszielen, als auch bei der Suche von geeigneten Kandidaten für Medikamentenanwendungen und somit in der Beschleunigung klinischer Studien eingesetzt. Ein weiterer wichtiger Punkt ist das Finden von Biomarkern von Krankheiten, als auch die

Behandlung zu personalisieren und herauszufinden, welche Merkmale des Patienten darauf hindeuteten, dass ein Patient auf eine bestimmte Behandlung anspricht. (Schmitt, 2019)

Algorithmus des maschinellen Lernens in dieser Doktorarbeit

Für den Begriff ‚Algorithmus' wird in dieser Doktorarbeit das Modell des RetinaNet verwendet. Dieses wird im folgenden Absatz kurz erläutert. Ergänzend wird das *ResNet*, das zur Screening-Untersuchung des C1 Datensatzes in der Arbeit von SCHOBER und somit zur Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial eingesetzt wird, kurz erklärt.

RetinaNet:

Das Model, das in dieser Arbeit stellvertretend für ‚Algorithmus des maschinellen Lernens' steht, ist das RetinaNet. LIN ET AL. führten dieses in Detektionsaufgaben ein (Lin, Goyal, Girshick, He, & Dollar, 2018). Das RetinaNet prognostiziert keine pixelgenauen präzisen Segmentierungen, sondern verwendet Begrenzungsrahmen (engl. boundingboxes) für lokalisierte Objekte. Ursprünglich basierten diese Vorhersagen auf zweistufigen Detektoren, wie Region Based Convolutional Neural Networks (R-CNN) (Girshick, Donahue, Darrell, & Malik, 2014). Diese fanden im ersten Schritt geeignete Objektpositionen und klassifizierten anschließend die vorgeschlagenen Begrenzungsrahmen. Das RetinaNet kombiniert diese beiden Schritte und wurde damit ein schneller, einstufiger Detektor mit einer hohen Detektionsgenauigkeit. Das Problem, das frühere einstufige Detektoren ziemlich ineffizient machte, war die damit verbundene Inferenzzeit. Die Inferenzzeit beschreibt die Zeit, die das System für das Treffen einer Entscheidung benötigt. (Lin et al., 2018)

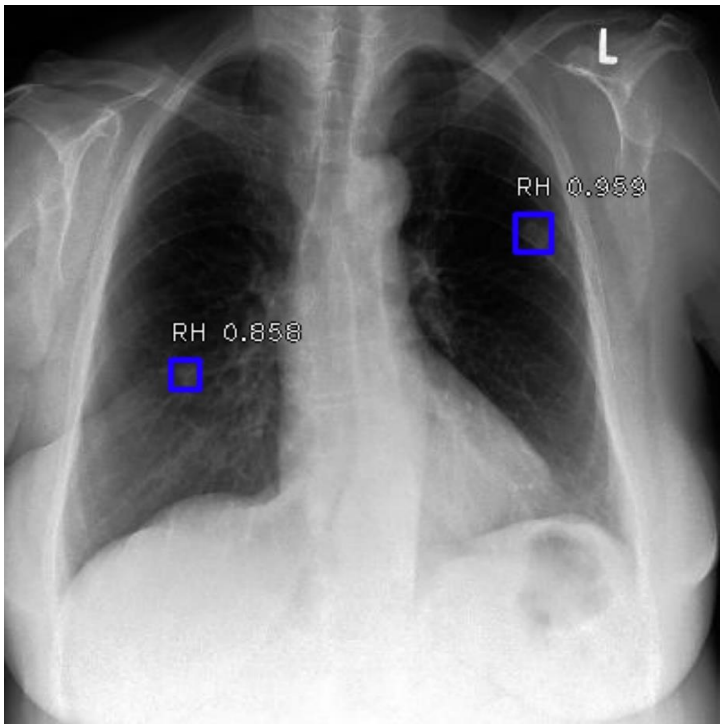


Abbildung 7: Vorhersagebeispiel des RetinaNet

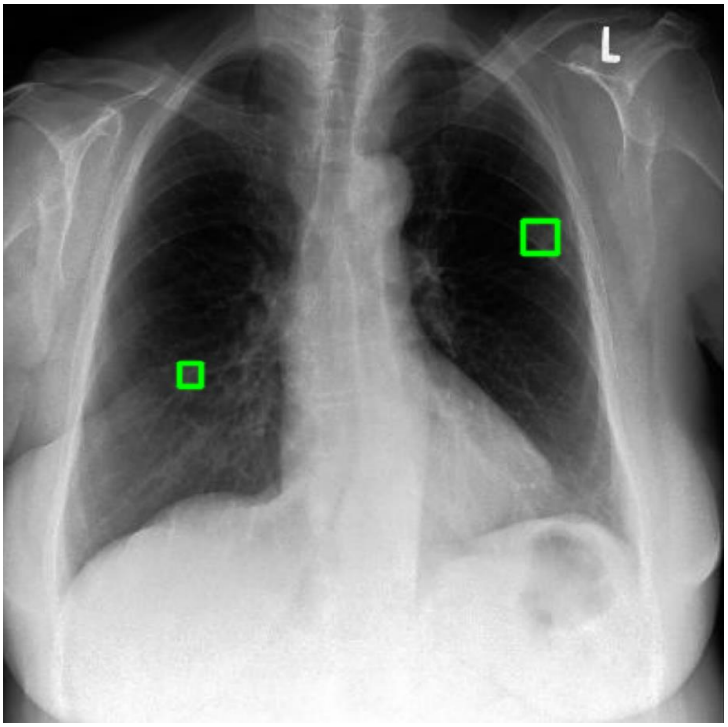


Abbildung 8: Groundtruth-Darstellung zu Abbildung 7

In dieser Arbeit wurde das RetinaNet Model für die Aufgabe der Erkennung von Lungenrundherden trainiert und anschließend mit der Leistung von Radiologen verglichen. Hiermit soll ein potentieller Einsatz im klinischen Alltag bewertet werden. In dieser Hinsicht ist es von essentieller Bedeutung, dass dieses sich gegenüber Fremdkörpern auf Thoraxröntgenbildern robust verhält. Das Verhalten des RetinaNet Models wurde untersucht und die Leistung mit den Teilnehmern der Reader Studie verglichen.

ResNet:

ResNet ist eine Kurzschreibweise für *Residual Neural Networks* und ist eine weitere Technik des *Convolutional Neural Networks* (Abk. CNN), die es ermöglicht Netzwerke erfolgreich zu trainieren, die sich wesentlich tiefer als die bisherigen CNN zeigen. Die Tiefe wird beschrieben durch die Anzahl von (Faltungs-)Schichten (engl. convolutional layer). Diese Technik hat ihren Ansatz in der Biologie, wo das Verhalten bestimmter Nervenzellen im menschlichen Gehirn (sogenannte Pyramidenzellen) nachgestellt wird, indem Abkürzungen zwischen den verschiedenen Schichten vorgenommen werden (engl. short-cut). (He, Zhang, Ren, & Sun, 2015) Ein CNN detektiert anhand von Filtern ortsunabhängige Strukturen aus Eingangsdaten (engl. input). Die Filter werden anhand von simplen Strukturen wie Linien, Kanten, Kontrasten und Farben aktiviert und vom Netz automatisch in diesem Prozess gelernt. Im darauffolgenden Schritt werden anschließend komplexere Strukturen, die aus Kombination der vorher genannten simplen Strukturen bestehen erlernt. Mit jedem weiteren Schritt, der auch als höhere Filterebene bezeichnet wird, kann sich somit das Abstraktionslevel des Netzes erhöhen. Aus den charakteristischen Merkmalen der vorhergehenden Schicht ergibt sich daraufhin die Aktivierung der letzten Faltungsschichten (engl. output). (Becker, 2019) Man fand heraus, dass mit einem Hinzufügen von weiteren Schichten die Performance der CNNs nicht unbedingt weiter ansteigt. ResNet führte nun sogenannte *Residual Connections* ein, welche die Performance, trotz weiterer Schichten, verbessern. (He et al., 2015) In dieser Arbeit wird das ResNet hinsichtlich des Screenings von Lungenrundherden bei gleichzeitigem Vorkommen von Fremdmaterial angewendet und die Ergebnisse in Kapitel 5 dargestellt.

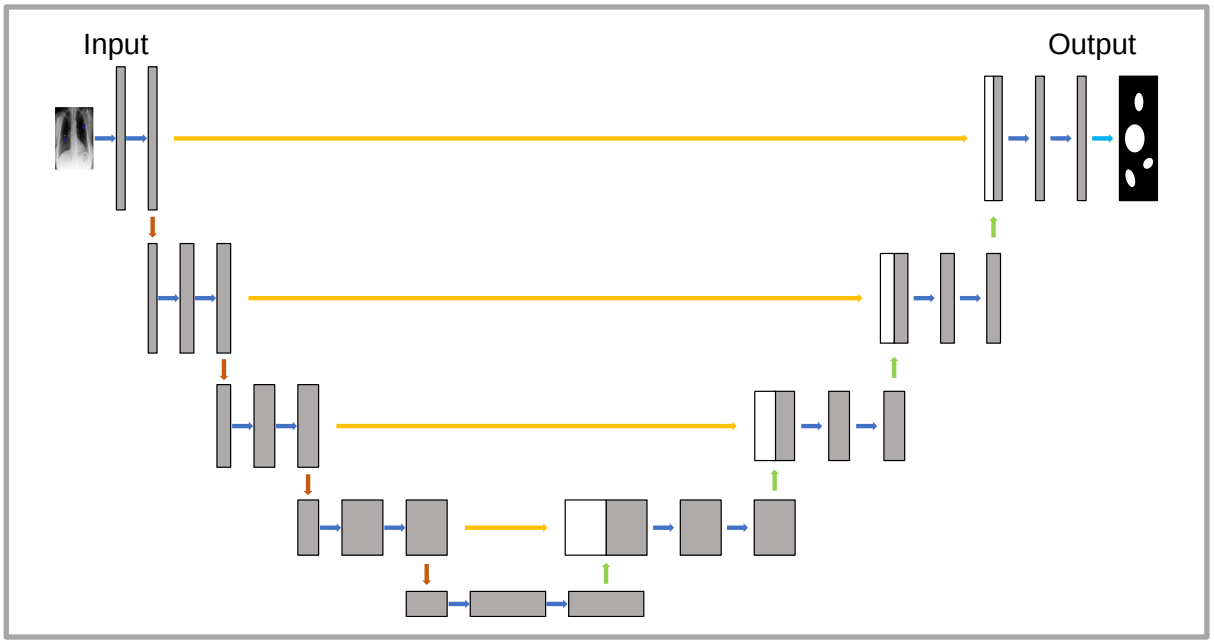


Abbildung 9: Exemplarischer Aufbau eines CNNs zur Segmentierung von Röntgenbildern in Anlehnung an (Ronneberger, Fischer, & Brox, 2015)

Die vom Betrachter aus linke Seite der Abbildung 9 zeigt die Input-Daten, wohingegen rechtsseitig die Output-Daten dargestellt sind. Weiße Kästen stellen kopierte Schichten mit charakteristischen Merkmalen dar, die als weitere Ausgangsbasis dienen. Die Pfeile kennzeichnen verschiedene CNN typische Operationen mit Pooling oder Konvolution. (Ronneberger et al., 2015)

2.4.3 Training und Unterteilung von maschinellem Lernen

Die Algorithmen des maschinellen Lernens, die für das Erkennen von Mustern und das Generieren von Lösungen verantwortlich sind, können durch verschiedene Lernkategorien trainiert werden.

Hierbei unterscheidet man überwachtes Lernen von teil- und unüberwachtem, sowie bestärkendes und aktives Lernen voneinander.

Das überwachte Lernen setzt voraus, dass der Algorithmus anhand von bekannten Daten und Beispielen lernt. Hierbei wird der Zusammenhang zu einer Zielvariablen erlernt und anschließend versucht diese richtig vorherzusagen. (Goram, 2018b; Tutanch, 2016) In der Radiologie kann die automatisierte Erkennung eines Lungenrundherds durch eine Röntgenaufnahme des Brustkorbs auch überwachtes Lernen bedeuten. In diesem Fall nähert sich der Computer dem an, was ein ausgebildeter Arzt bereits mit hoher Genauigkeit kann. Überwachtes Lernen wird häufig zur Abschätzung des Risikos eingesetzt. (Deo, 2015)

Dem überwachten Lernen steht das unüberwachte Lernen gegenüber, bei welchem die Modelgruppen selbstständig aufgrund eigenständig erkannter Muster gebildet werden. Ziel dieses Lernens ist die Ableitung von Regeln, welche von der Anzahl und der Art des Clusters der zugrundeliegenden Daten des Corpus abhängig sind. Das teilüberwachte Lernen ist eine Kombination aus dem überwachten und dem unüberwachten Lernen und wird für die gleichen Zwecke, wie das überwachte Lernen eingesetzt. Das bestärkende Lernen ist eine Sonderform des maschinellen Lernens und ist dem menschlichen Lernen sehr ähnlich. Die Aktivität des Algorithmus, wie er auf bestimmte Ereignisse zu reagieren hat, basiert auf Belohnungen und Bestrafungen. Beim aktiven Lernen wird zuvor eine Auswahl relevanter Fragen vorgenommen, die eine hohe Ergebnisrelevanz erzielen (Trabold, 2021). Diese Auswahl wird durch den Algorithmus selbst getroffen. Somit besteht die Möglichkeit für bestimmte Eingangsdaten die gewünschten Ergebnisse anzufragen. (Goram, 2018b; Tutanch, 2016)

Eine Untergruppe des maschinellen Lernens stellt das *Deep Learning* dar. Diese Methodik ist ähnlich strukturiert wie die Verarbeitung im menschlichen Gehirn. Deep Learning arbeitet mit künstlichen neuronalen Netzen, die zwischen der Eingabeschicht und der Ausgabeschicht zahlreiche Zwischenschichten herausbilden. Jede Zwischenschicht wird durch unterschiedliche Filter ausgewertet und daraufhin weiterverarbeitet, sodass jede Bewertung in einer anderen Ebene erfolgt und auf der Ausgabe der vorherigen Ebene basiert. Die Zwischenschichten werden auch versteckte Schichten

(engl. hidden layers) genannt und jeder verwendetet Filter pro Schicht erzeugt einen Output-Punktestand, der die Eingangsergebnisse der nächsten Ebene darstellt bis ein Endergebnis, wie beispielsweise eine Diagnose erreicht ist. (Mintz, 2019)

2.4.4 Computer-Aided Detection, Computer-Aided Diagnosis

Im folgenden Absatz werden die Begriffe Computer-aided-detection (CADe) und Computer-aided-diagnosis (CADx), als auch deren Funktionsweise kurz erläutert.
Computer-aided-detection wird seit dem Jahr 1966 anhand von Thorax- und Brustbildgebung entwickelt und seit der Jahrtausendwende klinisch eingesetzt. Voraussetzungen für den derzeitigen Stand der Entwicklung und die Validität des Programms ist die Verfügbarkeit eines großen Trainingdatensatzes, den Zugriff auf ausreichend große Rechenressourcen, einschließlich Computerleistung und Speicherung, sowie eine qualitativ hochwertige digitale Bildgebung. Ebenfalls von Bedeutung ist die Möglichkeit des elektronischen Zugangs auf die Informationen die zum Zeitpunkt der Datenerhebung relevant waren. Dazu gehört beispielsweise die Krankengeschichte, Untersuchungen oder Laborberichte des Patienten. Das System der CADe weist auf suspekte Regionen in der Bildgebung hin, die durch parenchymale Verdichtungen und Muster auffallen, wie beispielsweise beim Mammographie-Screening. Somit dient dieses Computersystem dem Radiologen als Unterstützung. Das CADe schlägt jedoch keine Diagnose vor und lässt somit die Diagnosestellung und die Verantwortung auf Seiten des Radiologen bestehen. (Giger, 2018)
CADx umfasst die Charakterisierung einer Region oder eines Tumors, die ursprünglich entweder von einem Radiologen oder einem Computer angezeigt wurde. Daraufhin charakterisiert der Computer die verdächtige Region oder Läsion und schätzt die Wahrscheinlichkeit einer Erkrankung ab. Die endgültige Diagnosestellung und der weitere Umgang mit dem Patienten liegt jedoch auf Seiten des Arztes. (Giger, Karssemeijer, & Schnabel, 2013) Abbildung 10 zeigt eine Übersicht über die Grundarchitektur von CAD-Systemen als Flussdiagramm. Hieraus geht hervor, dass nach der Segmentierung der Bilddatensätze, beispielsweise durch Grauwertverfahren, relevante Merkmale entweder manuell oder (semi-)automatisch extrahiert werden können. Anschließend werden diese durch statistische Verfahren, oder neuronale Netze klassifiziert und es entsteht ein Ergebnis, dass dem Radiologen zur Befundung und Diagnosestellung dient.

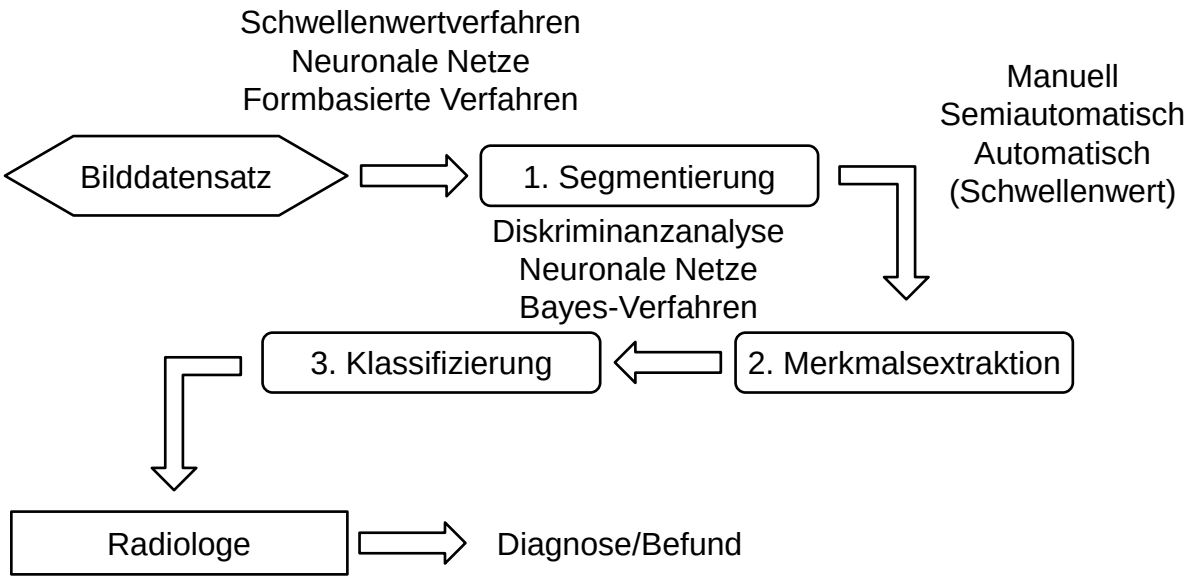


Abbildung 10: Grundarchitektur von CAD-Systemen in Anlehnung an (Achenbach, Vomweg, Heussel, Thelen, & Kauczo, 2003)

2.5 Grundlegende Parameter

Um die diagnostische Genauigkeit und Güte von Algorithmen des maschinellen Lernens zu bewerten, gibt es verschiedene Parameter, die Genauigkeit und Güte in Zahlen ausdrücken und somit zu messbaren Merkmalen machen können. Dazu gehören die Parameter der Vierfeldertafel, als auch der F2 Score und die Darstellung mittels ROC-Kurve, welche nachfolgend genauer erläutert werden.

2.5.1 Parameter der Vierfeldertafel

Wie zuvor bereits erwähnt, kann die Vierfeldertafel dazu verwendet werden die diagnostische Genauigkeit und Güte von Algorithmen des maschinellen Lernens zu bewerten. Zu den quantitativen Grundparametern der Vierfeldertafel, wie in Tabelle 6 dargestellt, zählen richtig positiv (Abk. TP), falsch positiv (Abk. FP), falsch negativ (Abk. FN) und richtig negativ (Abk. TN). Basierend auf diesen Grundparametern und den zugrundeliegenden Sachverhalten können weiterführende Berechnungen dazu verwendet werden, Aussagen über die Genauigkeit, Sensitivität und Spezifität des Algorithmus zu treffen.

TP beschreibt, dass das positive Testergebnis mit dem Vorliegen der Krankheit übereinstimmt, wohingegen bei FP das Testergebnis positiv ist, jedoch keine Krankheit vorliegt. FN sagt aus, dass der Test ein negatives Ergebnis liefert, obwohl die Krankheit eigentlich vorhanden ist. TN hingegen gibt ein negatives Testergebnis an, welches korrekterweise den Zustand als gesund identifiziert.

Tabelle 6: Übersicht über die Vierfeldertafel

		Krankheit liegt vor		
		Ja = Krank	Nein = Gesund	
Test	positiv	A TP	B FP	$A+B$ Alle Testpositiven
	negativ	C FN	D TN	$C+D$ Alle Testnegativen
		$A+C$ Alle Kranken	$B+D$ Alle Gesunden	$A+B+C+D$ Alle Analysen
		Sensitivität $\frac{A}{A+C}$	Spezifität $\frac{D}{D+B}$	

Das Maß der Genauigkeit setzt sich zusammen aus der Anzahl der korrekten Klassifizierungen dividiert durch die Anzahl der gesamthaft zu testenden Objekte.

$$\text{Genauigkeit} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Die Sensitivität setzt sich aus der Anzahl richtig positiver Ergebnisse dividiert durch die Summe von richtig positiven und falsch negativen Werten zusammen. Sie gibt an mit welchem Prozentsatz eine Erkrankung tatsächlich erkannt wird und wird somit auch als Richtig-positiv-Rate (TPR) bezeichnet.

$$\text{TPR, Sensitivität, Recall} = \frac{(TP)}{(TP+FN)}$$

Die Spezifität, auch als Richtig-negativ-Rate (TNR) bezeichnet, errechnet sich aus dem Quotienten von richtig negativen Testergebnissen und der Summe derer, denen tatsächlich keine Erkrankung zugrunde liegt. Somit gibt diese die Wahrscheinlichkeit an, mit der tatsächlich Gesunde, die nicht an der untersuchten Krankheit leiden, im Test auch als gesund erkannt werden.

$$\text{TNR, Spezifität} = \frac{(TN)}{(FP+TN)}$$

Die Falsch-positiv Rate-(FPR) beschreibt den Anteil der fälschlicherweise als erkrankt klassifizierten Personen, die eigentlich gesund sind. Hiermit wird die Wahrscheinlichkeit für einen Fehlalarm angegeben, da ein gesunder Patient zu Unrecht als krank diagnostiziert wird.

$$\text{FPR} = \frac{(FP)}{(FP+TN)}$$

Die Falsch-negativ-Rate (FNR) beschreibt den Anteil der Personen, die tatsächlich erkrankt sind, jedoch aber als gesund diagnostiziert wurden. Sie wird berechnet indem man die fälschlicherweise als negativ klassifizierten Patienten zu der Gesamtheit der tatsächlich erkrankten Patienten ins Verhältnis setzt.

$$\mathrm{FNR} = \frac{(FN)}{(FN + TP)}$$

2.5.2 F_β und F2 Score

Der F_β Score ist ein Maß zur Beurteilung der Güte von Algorithmen des maschinellen Lernens mit einer einzigen Kennzahl. Hierbei kombiniert der Score die bereits vorab erläuterten Parameter Genauigkeit (engl. precision) und Sensitivität (engl. recall) mittels des gewichteten harmonischen Mittels *β.* (Minto, 2009) Die unten stehende Formel stellt die Basis zur Berechnung des Scores da.

Formel 1: Formel zu Ermittlung von F_β, in Anlehnung an (Schober, 2019)

$$\boldsymbol{F_\beta = \frac{(1 + \beta^2) \cdot TP}{(1 + \beta^2) \cdot TP + \beta^2 \cdot FN + FP}}$$

Die Variable β bestimmt zu welchen Teilen die Genauigkeit und Sensitivität in das Ergebnis des F_β Scores einfließen. Wird β durch 1 ersetzt, so kann der F1 Score berechnet werden. Dieser wird häufig für grundlegende Klassifizierungsaufgaben verwendet. Bei medizinischen Anwendungen wird der Genauigkeit ein höherer Stellenwert zugeordnet, weshalb sich der F2 Score als repräsentativer Messwert etabliert hat. (Schober, 2019)

Um den F2 Score zu ermitteln muss β gleich 2 gesetzt werden. Bei der Verwendung des F2 Scores gilt, dass die Sensitivität mehr gewichtet wird als die Präzision. Aus diesem Grund ist der F2 Score in bestimmten Anwendungsbereichen, bei denen es wichtiger ist, so viele positive Stichproben wie möglich korrekt zu klassifizieren, anstatt die Anzahl der korrekten Klassifizierungen zu maximieren, besser geeignet. (Minto, 2009)

2.5.3 Definition der ROC-Kurve

Die *Receiver Operator Characteristics* (Abk. ROC) sind die am häufigsten verwendeten Leistungsindikatoren zur Bewertung der Diagnose von CAD-Systemen (CADx) (Firmino et al., 2016). Die *Receiver-Operating Characteristic curve* (Abk. ROC curve, dt. ROC-Kurve) stellt den Zusammenhang zwischen der Anzahl der richtig positiven Werte unter den Erkrankten und den falsch positiven Werten unter den Gesunden dar. Diese ergibt sich unter der Betrachtung vieler Paare von Sensitivität (TPR) und Spezifität (FPR) und einem für den Test individuell festgelegten Schwellenwert (engl. Cut-off). Der Schwellenwert bestimmt die Einordnung von Kategorien in zwei verschiedene Bereiche. Liegen Werte über dem festgelegten Schwellenwert, werden Wahrscheinlichkeiten der einen Kategorie zugeordnet. Liegen sie unterhalb des Schwellenwerts, werden sie der anderen Kategorie zugeordnet. (Hilgers, Heussen, & Stanzel, 2019)

Die Generierung der ROC-Kurve erfolgt in einem Koordinatensystem, in welchem die Sensitivität (TPR) auf der y-Achse gegen 1-Spezifität (FPR) auf der x-Achse

aufgetragen und durch die Wertepaare (0,0) und (1,1) begrenzt ist (siehe Abbildung 11). Eine ideale ROC-Kurve verläuft zunächst senkrecht, also parallel zur bzw. auf der y-Achse. Dieser Verlauf bedeutet, dass die Sensitivität zu Beginn nahezu bei 100 % und die Fehlerquote zu dem Zeitpunkt bei 0 % liegt. Der darauffolgende Verlauf entlang der oberen Seite des Quadrats lässt erkennen, dass die FPR ansteigt. Wird eine ROC-Kurve als Treppenfunktion dargestellt, kann man davon ausgehen, dass diese direkt aus den Daten ermittelt wurde. Verläuft die ROC-Kurve nahe der Winkelhalbierenden des Koordinatensystems, so bedeutet dies eine gleiche Anzahl von richtig und falsch positiven Werten, was einer Verteilung im Sinne eines Zufallsprozesses entspricht. Die Darstellung mittels ROC-Kurve zeigt besondere Relevanz wenn es um den Vergleich mehrerer diagnostischer Tests geht. Liegt eine ROC-Kurve vollständig oberhalb einer anderen ROC-Kurve, so ist diese im Gesamtmaß überlegen und zeigt eine bessere Leistungsfähigkeit. (Hilgers et al., 2019)

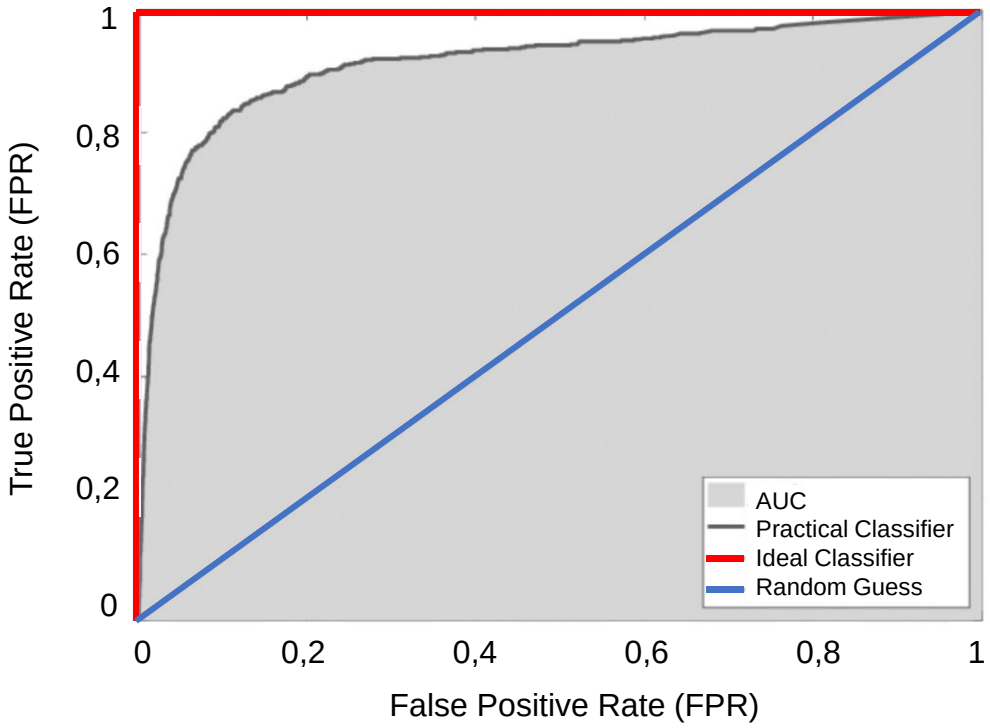


Abbildung 11: Prinzip der ROC-Kurve in Anlehnung an (Chen, 2019)

Die ROC-Kurve stellt einen Index dar, der sich auf die Genauigkeit und Effektivität bezieht, und zwar durch die Beziehung zwischen der Wahrscheinlichkeit wahrer und der Wahrscheinlichkeit falsch positiver Ergebnisse (Firmino et al., 2016). Somit ist eine ROC-Kurve eine Wahrscheinlichkeitskurve (Narkhede, 2019). Der allgemeine Verlauf einer ROC-Kurve ist invariant bezüglich streng monoton wachsender Transformationen, was bedeutet, dass die Wertepaare unabhängig voneinander einen steigenden Charakter aufweisen (Mansmann, 2011).

Um die Güte eines Algorithmus des maschinellen Lernens zu quantifizieren, kann die Fläche unter der ROC-Kurve berechnet werden. Im Englischen wird dieses als *Area*

Under the Curve (Abk. AUC) bezeichnet. Die Fläche kann dabei Werte zwischen 0 und 1 annehmen. Bei einem Wert von 0,5 entspricht dies einer zufälligen Vorhersage durch den Algorithmus des maschinellen Lernens. Nimmt die AUC einen Wert von 1 an, entspricht das einer perfekten Vorhersage durch den Algorithmus. (Chen, 2019)
Demnach repräsentiert die AUC das Maß der Prognose- oder Klassifizierungsfähigkeit des Modells und je höher ihr Wert ist, desto besser kann der Algorithmus eine richtige Prognose treffen. Hat eine AUC beispielsweise einen Wert von 0,6 so bedeutet dies, dass der Algorithmus mit einer Wahrscheinlichkeit von 60 % zwischen positiv und negativ unterscheiden kann. (Narkhede, 2019)

3 Bestehende Ansätze in der Forschung

Im nachfolgenden Absatz werden einige bestehenden Ansätze in der Forschung des maschinellen Lernens vorgestellt. Ein besonderer Fokus liegt hierbei auf der Erstellung realitätsnaher Datensätze und der Detektion von Lungenrundherden im Rahmen der diagnostischen Bildgebung. Anhand der existierenden Ansätze wird anschließend die Relevanz dieser Doktorarbeit herausgearbeitet.

Im Laufe der letzten Jahre erlangte die Forschung über die Detektion von Lungenrundherden mittels maschinellem Lernen stetig mehr an Bedeutung. Basis der Forschung, bezogen auf das Screening und die Detektion von Lungenrundherden, bildet derzeit die Verwendung von CT- und Thoraxradiographie-Bilddatensätzen. Für die vorliegende Forschungsarbeit sind saubere und realitätsnahe Datensätze mit unterschiedlichen Eigenschaften, wie beispielsweise das Vorliegen von Rundherden, Fremdmaterial, sichtbaren Erkrankungen auf dem Thoraxröntgenbild oder gesunde Thoraxröntgenbilder essentiell für Entwicklung, Training und Testung von Algorithmen des maschinellen Lernens. Derzeit gibt es nahezu keine gesonderten Ansätze in der Literatur, die auf den ersten Teil der Arbeit, die Erstellung von geeigneten Datensätzen zur Entwicklung von maschinellem Lernen, fokussiert sind. Deutlich mehr Literatur gibt es zum zweiten Teil dieser Arbeit, die sich auf die Bewertung der Detektionsleistung von Algorithmen des maschinellen Lernens bezieht. Ansätze die sowohl für den ersten, als auch für den zweiten Teil dieser Arbeit von Relevanz sind, werden nachfolgend kurz erläutert.

SHIRAISHI ET AL. entwickelten eine digitale Bilddatenbank mit 247 Thoraxröntgenbildern. Das Ziel dieser Studie war es, die Eigenschaften von Bilddatenbanken für den möglichen Einsatz in verschiedenen digitalen Bildforschungsprojekten zu untersuchen. 154 konventionelle Thoraxröntgenbilder mit Lungenrundherden und 93 konventionelle Thoraxröntgenbilder ohne Lungenrundherde wurden aus 14 medizinischen Zentren ausgewählt und mit einer Größe von 2048x2048 Pixeln digitalisiert. Die vorhandenen Lungenrundherde wurden in fünf Subtilitätsgrade, von offensichtlich bis sehr unscheinbar, eingeteilt. Die Beobachtungen von 20 teilnehmenden Radiologen wurden mittels ROC-Kurve dargestellt. Durch diese Darstellung konnte die Detektionsrate hinsichtlich der vorliegenden Lungenrundherde verglichen werden. Die Analyse der ROC-Kurve und die AUC mit Werten von 0,547 bis 0,991 zeigte, dass die Datenbank eine Variation von Lungenrundherden mit unterschiedlichen Eigenschaften enthält und somit für viele Zwecke beispielsweise in der Forschung, Bildung, Qualität, Versicherung und Demonstration von Nutzen sein kann. (Shiraishi et al., 2000b)

Darüber hinaus forschte **SCHMETTE** an kombinierten Ansätzen für die Segmentierung von Lungentumoren mit Hilfe von konventionellen Thoraxröntgenbildern und Dunkelfeldröntgenbildern. In SCHMETTES Arbeit aus 2018 wurde ein Lungensegmentierungs-

netzwerk verwendet, das aus gesunden CT-Daten Lungentumore generierte und diese sowohl in Thoraxröntgenaufnahmen, als auch in Dunkelfeldröntgenbilder durch Vorwärtsprojektion einarbeitete. In einer vergleichenden Reader Studie wurde ein Algorithmus auf Basis von Radiographie- und Dunkelfelddaten trainiert, um die Tumore zu segmentieren. Die Untersuchung zeigte, dass die Verwendung zusätzlicher Dunkelfelddaten bei der Detektion von Lungenrundherden den Nachweis dieser im Tumorscreening verbessern könnte. Hierbei wird insbesondere die Anzahl falsch positiver Ergebnisse reduziert. Zudem zeigte der Algorithmus im Vergleich zu den Radiologen ein besseres Ergebnis in der Detektion der Lungentumore. Jedoch gilt es zu berücksichtigen, dass für die Generierung eines Lungentumors und dessen Position im Thoraxröntgenbild lediglich ein erweitertes CT-Bild eines Tumors verwendet wurde. Dies bedeutet, dass in dieser Studie derselbe Tumor für das Training, die Validierung und den Test verwendet wurde, was wiederum die hohe Variabilität von Tumoren in realen Patientendaten nicht reproduziert. (Schmette, 2018)

NAM ET AL. entwickelten ein DLAD-Netzwerk (Deep Learning Automation Detection), welches maligne Lungenrundherde auf Thoraxröntgenbildern erkennen sollte. Für das Training und die Erstvalidierung wurde ein Datensatz des Seoul National University Hospital verwendet. Hierbei wurden 43292 posterior-anteriore Thoraxröntgenbilder eingeschlossen, von denen 3213 Bilder suspekte, maligne Läsionen aufwiesen. Das Netzwerk und die Bearbeitung der Datensätze wurden halbüberwacht. Die Auswertung verschiedener Datensätze unterschiedlicher Krankenhäuser zeigte, dass die AUC für das Screening der Lungenrundherde zwischen 0,92 und 0,99 lag. 18 Radiologen mit unterschiedlichen Erfahrungsstufen wurden für eine Reader Studie rekrutiert, um ihre Leistung mit dem Algorithmus zu vergleichen. Der DLAD Ansatz übertraf die meisten Radiologen sowohl beim Screening als auch beim Nachweis der Lungenrundherde. Zogen Radiologen jedoch den Algorithmus als Unterstützung hinzu, so zeigten diese eine verbesserte Leistung bei der Detektion maligner Lungenrundherde. In dieser Untersuchung wurden keine weiteren Lungenerkrankung berücksichtigt und auch die Unterscheidung zwischen maligne und benigne wurde vernachlässigt. Zudem gibt es keine genauen Daten zur Detektionsrate von Lungenrundherden, die kleiner als einen Zentimeter Durchmesser aufwiesen. (Nam et al., 2019)

Zudem trainierten und validierten **RAJPURKAR ET AL.** einen Algorithmus namens „CheXNeXt“ (eine verbesserte Version des ChestX-ray8-Datensatzes). Der interne Validierungssatz bestand aus 420 Bildern, von denen mindestens 50 Bilder 10 verschiedene Pathologien wie Pneumonie, Pleuraerguss oder Lungenrundherde in anterior-posterior bzw. posterior-anterior Aufnahmen enthielten. Neun ausgebildete Radiologen wurden als Referenz für den Vergleich verwendet. Durchschnittlich benötigten die Radiologen 240 Minuten für 420 Thoraxröntgenbilder. Verglichen hierzu benötigte der Algorithmus nur 1,5 Minuten. Die Grundwahrheit wurde durch drei gleiche Ergebnisse von Radiologen, die auf Thoraxröntgenbilder spezialisiert sind, gebildet. In

Bezug auf das Screening von Lungenrundherden wies der Algorithmus eine AUC von 0,900 auf. Im Vergleich dazu, zeigten die Radiologen eine durchschnittliche AUC von 0,933 und übertrafen somit das Ergebnis des Algorithmus. In drei von zehn Pathologien im Thoraxröntgenbild zeigten die Radiologen bessere Ergebnisse als der Algorithmus. Dennoch zeigte der Algorithmus beim Vorliegen von Atelektasen ein besseres Ergebnis als die Radiologen. (Rajpurkar et al., 2018)

MARTEN ET AL. zeigte anhand einer prospektiv verblindeten Beobachtungsstudie den Vergleich eines interaktiven Prototypens des CAD-Systems und zwei erfahrenen Radiologen für die Detektion von Lungenrundherden im MSCT (Multislice CT). Hierbei wurde besonderer Fokus auf den Einfluss von Größe, Position, Morphologie, Gefäß- und Pleuranähe der Lungenrundherde gelegt. Bei 20 Patientenbildern und somit insgesamt 135 Lungenrundherden betrug die Detektionsrate des CAD-Systems 76,3 % und der Radiologen jeweils 52,6 %. Die FPR des CAD-Systems betrug 0,55, wohingegen die Radiologen eine FPR von 0,25 und 0,15 zeigten. Koppelte man die Detektionsleistung des CAD-Systems und die der Radiologen, so stieg die Detektionsrate auf 93,3 % und die FPR fiel auf 0,1 pro Aufnahme. Das Ergebnis dieser Untersuchung zeigte, dass das CAD-System die FPR senkt, die Detektion kleiner Lungenrundherde ohne Gefäßkontakt sehr sensitiv detektiert und somit die Leistung der Radiologen kompensieren kann und damit die Detektionsrate allgemein erhöht. (Marten et al., 2005)

Bereits 2005 verglichen **RUBIN ET AL.** die Leistung von Radiologen mit der eines durch maschinelles Lernen trainierten Algorithmus anhand der Detektion von Lungenrundherden in 20 dünnschichtigen (1,25 mm) CT-Aufnahmen. Die Referenz bildeten vorab zwei erfahrene Thoraxradiologen, wohingegen drei weitere Radiologen unabhängig voneinander die CT-Aufnahmen analysierten und die Position des Lungenrundherds sowie einen Konfidenzwert festlegten. Das CAD-System wurde mithilfe von Kreuzvalidierung ausgewählt und ebenfalls auf die 20 CT-Aufnahmen angewandt. Die Ergebnisse wurden anhand TP, FP, Konfidenzniveau und ROC-Kurven festgelegt. Die Leistung der Befundung wurde auf der Grundlage von TP Detektionen beider Radiologen ermittelt. 20 CT-Aufnahmen zeigten 195 nicht kalzifizierte Lungenrundherde mit einem Mindestdurchmesser von 3 mm. Die AUC der ROC-Kurven der drei Radiologen betrug 0,54, 0,48, und 0,55. Das CAD-System zeigte eine AUC von 0,36. Die Unterschiede zwischen Radiologe 3 und dem CAD-System, als auch zwischen Radiologe 2 und 3 waren signifikant ($p < 0,05$). Kein signifikanter Unterschied wurde hingegen zwischen dem CAD-System und Radiologe 1, als auch 2, gefunden. Die mittlere Sensitivität der Radiologen betrug 50 % (41–60 %). Die Doppelbefundung führte zu einem Anstieg auf 63 % (56–67 %). Als das CAD-System bei einem Schwellenwert verwendet wurde, der nur drei FP Detektionen pro CT-Aufnahme zuließ, wurde die Sensitivität auf 76 % erhöht (73–78 %). RUBIN ET. AL zeigte, dass das CAD-System mit einem Schwellenwert von bis zu 3 FP Detektionen, als Ergänzung zu einem Radiologen, Lungenrundherde

effektiver detektiert, als es in der Doppelbefundung durch zwei Radiologen der Fall war. Somit können CAD-Systeme nach RUBIN ET. AL die Leistung der Radiologen unterstützen, jedoch diese nicht ersetzen. (Rubin et al., 2005)

BLEY ET AL. evaluierten in 2008 die Leistung von CAD-Systemen im Vergleich zu Radiologen anhand der Detektion von subtilen Lungenrundherden mit einer Größe von 5–15 mm. Der Referenzstandard wurde durch zwei erfahrene Radiologen und das Vorliegen von CT-Aufnahmen festgelegt. Innerhalb von vier Wochen wurden zudem weitere Thoraxröntgenaufnahmen der Patienten angefertigt. Ausschließlich posterior-anteriore Aufnahmen wurden zur Evaluation des CAD-Systems und der Radiologen analysiert. Von 117 Patientendaten wurden 42 Patienten mit 66 Lungenrundherden mit mittlerem Lungenrundherd-Durchmesser von 7,5 mm (Standardabweichung 2,2 mm) in die statistische Analyse einbezogen. Die übrigen 75 Patienten zeigten keine Lungenrundherde in der Größe von 5–15 mm. Das CAD-System wies eine Sensitivität von 39,9 % und eine Detektion von 2,5 FP pro Bild auf und zeigte somit eine höhere Nachweisempfindlichkeit als die Radiologen. Diese wiesen eine Sensitivität zwischen 18,2 % und 30,3 % auf. Der Vergleich zwischen Radiologen und CAD-System zeigte nur eine mäßige Übereinstimmung, somit wurde das CAD-System hinsichtlich der Ergebnisse als Unterstützung für die Radiologen als sinnvoll erachtet. (Bley et al., 2008)

BUSH untersuchte den Einsatz eines ResNet CNN-Modells mit der Eigenschaft des Transfer-Lernens, um in Thoraxröntgenbildern das Vorkommen und die Lokalisation von Lungenrundherden, als auch deren Dignität zu ermitteln. Das Modell war in der Lage Lungenrundherde von keinen Lungenrundherden mit einer Sensitivität von 92 % und einer Spezifität von 86 % zu unterscheiden. Dennoch war es weniger in der Lage, zwischen gutartigen und bösartigen Lungenrundherden zu differenzieren. Ebenso konnte das Modell Regionen anzeigen, in denen sich der Lungenrundherd befand, jedoch weniger die genaue Position dessen. Der untersuchte Datensatz, welcher mittels CT-Aufnahmen, Gewebeproben und Verlaufskontrollen vorab verifiziert wurde, enthielt 93 Thoraxröntgenbilder ohne Lungenrundherde, 54 Thoraxröntgenbilder mit benignen Lungenrundherden und 100 Thoraxröntgenbilder mit malignen Lungenrundherden. Bei den verwendeten Thoraxröntgenbildern, die von einer japanischen Datenbank mit dem Namen *Japanese Society of Radiological Technology* (Abk. JSRT) stammten, handelte es sich um posterior-anterior Aufnahmen mit einer Größe von 2048x2048 Pixeln. Diese Bilder enthielten Informationen wie die Größe, als auch die Koordinaten des Lungenrundherds. Mit diesen Informationen konnte somit die Position der Begrenzungsrahmen angenähert und mit der tatsächlichen Position verglichen werden. Um den Datensatz fürs Training zu vergrößern, wurde die Anzahl der Trainingsbilder mittels Augmentationstechnik erhöht. (Bush, 2016)

Vor dem Hintergrund dieser jüngsten Erkenntnisse liegt der Fokus dieser Doktorarbeit auf der Erstellung geeigneter Datensätze, die vielseitig einsetzbar sind. Diese Datensätze können sowohl zum Training als auch zur Bewertung und zum Vergleich

moderner Techniken des maschinellen Lernens eingesetzt werden. Hierbei liegt ein besonderer Fokus auf der Auswertung in der konventionellen Radiographie. Betrachtet man die derzeitigen Forschungsergebnisse hinsichtlich Screening und Detektion von Lungenrundherden anhand von Algorithmen des maschinellen Lernens, so ist erkennbar, dass diese häufig auf der Basis von CT-Aufnahmen, jedoch nicht auf der Basis der konventionellen Radiographie erstellt wurden.

Die klinische Anwendung von Algorithmen des maschinellen Lernes auf Thoraxröntgenbilder birgt im Vergleich zur Computertomographie Vorteile. Zu nennen ist beispielsweise die reduzierte Strahlenbelastung für den Patienten (Alkadhi et al., 2012), welche wiederum einen verkürzten Screening-Zyklus ermöglicht. Zudem ist eine nahezu flächendeckende Verfügbarkeit der Gerätschaften, als auch ein leichteres Auswertungsprofil gewährleistet. (Ngoya, Muhogora, & Pitcher, 2016a) Somit können Verläufe durch eine untersucherunabhängige Beobachtung objektiviert werden (Dreher, Berthold, Nilius, Woehrle, & Rembert, 2019). Darüberhinaus sind Thoraxröntgenaufnahmen die häufigste medizinische Bildgebung der Welt und somit entscheidend für die Diagnose vieler Erkrankungen im Bereich des Thorax (Rajpurkar et al., 2018).

Ziel dieser Doktorarbeit ist es, die praktische Anwendung von Algorithmen des maschinellen Lernens auf konventionelle Thoraxröntgenbilder in einem klinischen Rahmen zu bewerten. Hierbei wird die Leistung des Algorithmus des maschinellen Lernens bezogen auf die Detektion von Lungenrundherden untersucht. In diesem Zusammenhang wird vorab ein realitätsgetreuer Datensatz erstellt. Anschließend wird die Leistung von Algorithmen des maschinellen Lernens und zweier Radiologen in Bezug auf die Detektion von Lungenrundherden verglichen. Die händische Erstellung eines sauberen, realitätsgetreuen Datensatzes bildet die Basis zum Anlernen des Algorithmus, sowie zur Steigerung seiner anschließenden Detektionsleistung. Der Einsatz von Thoraxröntgenbildern mit verschiedenen Eigenschaften, analog den erstellten Unterkategorien in dieser Arbeit, stellt eine Besonderheit des Trainings und des Tests des Algorithmus dar. Somit ist dieser Datensatz nahezu ein Alleinstellungsmerkmal dieser Arbeit. Zudem zeigt diese Arbeit die Besonderheit der Analyse der Detektion von Lungenrundherden bei gleichzeitigem Vorkommen verschiedener Fremdmaterialien. Das Vorhandensein von aufliegendem oder implantiertem Fremdmaterial in Thoraxröntgenbildern ist keine Seltenheit in der klinischen Routine. Dennoch werden Algorithmen des maschinellen Lernens in diesem Bereich derzeit überwiegend mit Daten trainiert, die diesen Aspekt nicht berücksichtigen.

4 Konzeption der Methodik zur Klassifizierung von Thoraxröntgenbildern

Das nachfolgende vierte Kapitel stellt das Konzept zur systematischen Identifikation und Klassifikation von Thoraxröntgenbildern dar. Einleitend werden die technischen Voraussetzungen zur Erstellung und die Ausgangsbasis der Daten thematisiert. Darauf folgt die Erläuterung der Maßnahmen zur Datengewinnung, und die Definition der formalen Anforderungen an die zu erarbeitende Methodik. Diese sollen vor allem einer einheitlichen Beachtung und Verfolgung der zugrundeliegenden Kategorisierung dienen.

4.1 Technische Voraussetzung und Ausgangsbasis der Daten

Kapitel 4.1 erläutert die allgemeinen technischen Voraussetzungen zur Erstellung, Weiterverarbeitung und Archievierung von Patientendaten und dessen Bildgebungen. Die Aufnahme eines Patienten in eine ambulante oder stationäre Versorgung und die Anwendung von medizinischen Maßnahmen, unter anderem, die der bildgebenden Verfahren, erfordert die Aufzeichnung von Patientendaten. Die Patienten-Bild-Datenbank im Klinikum rechts der Isar arbeitet mit dem *Picture Archiving and Communication System* (Abk. PACS).

Dieses funktioniert auf der Basis digitaler Rechner und Netzwerke (Krüger-Brand, 2006). Das PACS hat eine Anbindung an die bilderzeugenden Modalitäten und ist die zentrale Komponente für ein digitales Bilddatenmanagement. Ursprünglich wurde es in der Radiologie etabliert, doch inzwischen ist es Bestandteil alle klinischen Disziplinen. Mittlerweile stehen die reibungslose Kommunikation und Vernetzung von Bildern und Befunden, sowie weiteren Daten und Dokumenten zwischen den Behandelnden im Mittelpunkt. (Mildenberger, 2011)

Zugleich ist das PACS mit Betrachtungs- und Nachbearbeitungsrechnern über einen krankenhausinternen Server verbunden. Darüber hinaus besteht eine weitere Verbindung zu einem System, das als *Radiologie-Informationssystem* (Abk. RIS) bezeichnet wird. Das RIS dient der radiologischen Abteilung zur Dokumentation und Verwaltung von medizinischen und administrativen Daten und basiert auf einem EDV-System. (Preisner, 2011) Etabliert wurde es ebenfalls vor ca. 20 Jahren in der radiologischen Abteilung. Heute dient es dem Arbeitsablauf aller an der radiologischen Bildgebung beteiligten Gruppen. Unterstützende Funktionen hat das RIS bereits bei der Anmeldung und Untersuchung des Patienten, als auch bei weiteren Aufgaben im Backoffice, wie Abrechnung bzw. Leistungsanalyse. (Mildenberger, 2011)

Als Voraussetzung all dieser Verknüpfungen miteinander ist ein weiteres System notwendig. Dieses System nennt sich *Digital Imaging and Communications in Medicine* (Abk. DICOM) und ist herstellerunabhängig mit allen Geräten kombinierbar. Es ist ein

System, dass die Integration der verschiedenen Komponenten zwischen dem PACS und RIS gewährleistet. DICOM standardisiert das Format der Daten, sowie das Kommunikationsprotokoll zu deren Austausch. Gleichzeitig gewährleistet es dadurch Interoperabilität zwischen Systemen verschiedener Hersteller. Ein System, dass die Historie des Patienten aus dem RIS mit der Befundung aktueller Bilder aus dem PACS vereint, nennt sich SECTRA IDS7. (Gebhardt, 2017) Dieses findet in der radiologischen Abteilung des Klinikums rechts der Isar Anwendung. SECTRA IDS7 bietet den Vorteil, dass die befundeten Patientendaten primär auf der Plattform des PACS gespeichert bleiben und der Radiologe direkt in seinem Befundtext Verweise und Verlinkungen auf Bilder und Historie des Patienten erstellen kann (Gebhardt, 2017).
Durch die Möglichkeit, Röntgenbilder digital zu speichern, wird ein Qualitätsverlust des Bildes auch nach länger bestehender Zeit verhindert. Im Sinne der Teleradiologie ist eine örtlich und zeitlich flexible Betrachtung der Diagnostik ermöglicht, die Gefahr des Verlustes, aufwändiges Archivieren und Transportkosten entfallen. Hervorzuheben ist somit die Steigerung von Effizienz, Funktionalität und Qualität im Vergleich zur analogen Bilddokumentation. (Akan, 2018)
Nach Genehmigung des Ethikantrages durch die entsprechende Ethikkommission im Februar 2018, in dem die Nutzung des PACS zugestimmt wurde, bestand die Möglichkeit auf bestehende Patientendaten zurückzugreifen. Durch die Größe der Datenbank und den Zugriff auf die gespeicherte Bildgebung ist es dem IT-System des Klinikums rechts der Isar möglich, anhand von Suchbegriffen entsprechende Patientendaten zu extrahieren.

4.2 Erstellung des Datensatzes

Dieser Abschnitt stellt den Prozess der Ermittlung der Daten durch geeignete Stichwortsuche und die Einordnung derer, um eine Basis für die Gestaltung des Konzepts zur systematischen Identifikation der Lungenrundherde, dar. Die Methodik der Gewinnung der Patientendaten und die numerische Zusammensetzung des Datensatzes sind im Folgenden beschrieben.

Stichwortsuche und vorläufige Einordnung der Daten

Für die Auswahl geeigneter Thoraxröntgenbilder wurde das PACS des Instituts für Radiologie nach entsprechenden Röntgenaufnahmen durchsucht. Die Patienten, die sowohl die spezifischen Stichwörter im Patientenbefund als auch Thoraxröntgenbilder in einer posterior-anterioren sowie seitlichen Aufnahme aufwiesen, wurden extrahiert. Diese Aufnahmen wurden alle im DICOM Format gespeichert. Der extrahierte Patientendatensatz wurde daraufhin manuell auf deren Befunde und Thoraxröntgenbilder hinsichtlich der Stichwörter überprüft. Extrahierte Patientendaten, die im gültigen

Befund oder Thoraxröntgenbild nicht mit den Stichwörtern übereinstimmten, wurden eliminiert.
Tabelle 7 zeigt die Anzahl von Patientendaten, die nach Prüfung auf Eignung verblieben sind und weiterverwendet wurden. Ebenso ist der Zeitraum, aus dem die Aufnahmen stammen, die zur Weiterverwendung genutzt wurden in der untenstehenden Tabelle dokumentiert.

Tabelle 7: Übersicht über die Stichwortsuche aus dem PACS

Stichwort	Anzahl Patienten	Anzahl bildgebender Untersuchungen	Zeitraum
Unauffälliger Herz-Lungen-Befund ab 2015	2098	2201	01.01.2015 – 03.03.2018
Rundherd	196	203	18.02.2004 – 31.01.2018
Lungenmetastase, Rundherd	63	66	12.02.2004 – 22.12.2017
Lungenmetastasen, Rundherde	144	182	13.01.2004 – 24.01.2018
Rundschatten, Metastase, Rundherd in Th2E	577	695	04.01.2010 – 26.03.2018

‚Unauffälliger Herz-Lungen-Befund' und ‚Rundherd' waren die ersten festgelegten Stichwörter. Patientendaten, dessen Befund diese Stichwörter enthält, wurden daraufhin automatisch zunächst in die Gruppen unauffälliger Herz-Lungen-Befund und Rundherd eingeordnet. Die Patientenaufnahmen, die sich nach Stichwortsuche und Sortierung in der Gruppe unauffälliger Herz-Lungen-Befund befanden, wurden vom 01.01.2015 bis zum 03.03.2018 aufgenommen. Nach Prüfung auf das Zutreffen dieser Eigenschaft zeigte sich eine Anzahl von 2098 Patienten. Bei der manuellen Überprüfung der Befunde mit dem zugehörigen Röntgenbild zeigte sich, dass die Patienten, die durch das Stichwort unauffälliger Herz-Lungen-Befund extrahiert wurden, erwartungsgemäß auch im wesentlichen unauffällige Befunde aufwiesen.
Der Begriff Rundherd hingegen ist um einiges unspezifischer und auch vermehrt mit vorangestellten Zusätzen versehen. Diese Zusätze, wie beispielsweise Verdacht auf, kein Verdacht auf, Hinweis auf, kein Hinweis auf, multiple, disseminierte, am ehesten, stehen häufig im Kontext mit dem Begriff Rundherd, sodass bei der Sortierung dieser Patienten der Befund und die Bildgebung eine ausschlaggebende Rolle spielte. Nach

Sortierung zeigte sich eine Anzahl von 196 Patienten aus dem Zeitraum 18.02.2004 bis 31.01.2018.
Um die Anzahl der Daten zu vergrößern, wurde im nächsten Schritt mit den Stichwörtern ‚Lungenmetastase + Rundherd', ‚Lungenmetastasen + Rundherde', sowie dem zusammenhängenden Suchbegriff ‚Rundschatten, Metastase und Rundherd in Th2E' gearbeitet. *Th2E* steht hier für Thoraxröntgenbild in 2 Ebenen (posterior-anterior und seitlich).
Die vorab genannten Stichwörter, wie Lungenmetastase + Rundherd erweiterten nach Sortierung den Patientendatensatz mit Aufnahmen aus dem Zeitraum 12.02.2004 bis 22.12.2017 um 63 Patienten. Der Plural dieser Suchbegriffe, entsprechend Lungenmetastasen + Rundherde, extrahierte im Zeitraum vom 13.01.2004 bis 24.01.2018 144 Patientendaten. 2016 Patienten wurden mit den Stichwörtern Rundschatten, Metastase, Rundherd in Th2E aus dem PACS herausgefiltert. Nach Sortierung dieser, verblieben 577 Patienten, deren Bilder im Zeitraum 04.01.2010 bis 26.03.2018 entstanden sind.
Da viele Patienten mehr als nur eine Thoraxröntgenbildgebung im Verlauf erhalten haben, war die Anzahl der Patienten häufig geringer als die Anzahl der bildgebenden Untersuchungen.
Die vorliegenden Daten, aufgeführt unter ‚Anzahl Patienten', ‚Anzahl bildgebender Untersuchungen', sowie der ‚betrachtete Zeitraum der Aufnahme' sind in Tabelle 7 dargestellt. Diese Daten wurden im Anschluss auf das Vorliegen von Rundherden, Nebenerkrankung und Fremdmaterial händisch genauer untersucht und anschließend nach den in Kapitel 4.3 beschriebenen formale Bedingungen zur kategorialen Einteilung und Aufbereitung des Datensatzes entsprechenden Kategorien zugeordnet.

4.3 Aufbereitung des Datensatzes

Im nachfolgenden Abschnitt werden die Ober- und Unterkategorien zur Klassifizierung der Patientendaten aus Kapitel 4.2 definiert. Hierauf aufbauend werden zudem die formalen Bedingungen zur Klassifizierung dieser erläutert. Eine Übersicht über die Einteilung des Datensatzes in die einzelnen Kategorien ist in Tabelle 8 aufgeführt. Thoraxröntgenbilder, die sich mit keiner Eindeutigkeit in einer dieser Kategorien zuordnen ließen, wurden aus dem Datensatz entfernt.

Formale Bedingungen zur kategorialen Einteilung

Im Rahmen der vorliegenden Dissertation wurden zunächst zwei Oberkategorien definiert, die sich im Vorkommen von Lungenrundherden im Thoraxröntgenbild unterscheiden. Lagen keine Lungenrundherde in der Bildgebung vor, so wurde diese Gruppe übergeordnet als ‚Unauffällig' bezeichnet. Waren jedoch eindeutig Lungenrundherde im Thoraxröntgenbild zu erkennen, so wurden diese der Kategorie ‚Rundherd'

zugeordnet. Bestand keine Eindeutigkeit bei der Zuordnung von Lungenrundherden im Thoraxröntgenbild, so wurde der Fokus auf diese Bilder gelegt, die im Verlauf der Untersuchung ein CT-Bild zur Abklärung erhielten. Das CT-Bild gab in den meisten Fällen Aufschluss über das Vorkommen und die Dignität und sicherte somit die korrekte Zuordnung in die Kategorien. Aufgrund der sich stark ähnelnden Patientenbefunde und Häufung ähnlicher Pathologien wurden die zwei Oberkategorien zur genaueren Klassifizierung in weitere Unterkategorien aufgeteilt. Hieraus ergaben sich somit insgesamt 7 Unterkategorien. Die vier Unterkategorien ‚Unauffällig' (Abk. U), ‚Unauffällig +'(Abk. U +), ‚Unauffällig + Mamillenschatten' (Abk. U + M) und ‚Unauffällig + Fremdmaterial' (Abk. U + FM), zählten zu der Oberkategorie Unauffällig. Der Oberkategorie Rundherd hingegen gehörten nur drei Unterkategorien an. Diese wurden als ‚Rundherd' (Abk. RH), ‚Rundherd +' (Abk. RH +) und ‚Rundherd + Fremdmaterial' (Abk. RH + FM) bezeichnet.

Die Kategorie, die als Unauffällig betitelt wurde, enthielt Thoraxröntgenbilder, die keine weiteren Auffälligkeiten aufwiesen. Das bedeutete, dass in dieser Kategorie keine Thoraxröntgenbilder mit Nebenerkrankungen oder Fremdmaterial im Thoraxbereich zu finden waren. Die Aufnahmen der Patienten entsprachen meist einem alterstypischen Herz-Lungen-Befund. Anatomische Strukturen, die sich im Thoraxröntgenbild etwas prominenter als normal erwiesen, aber keine Pathologie im eigentlichen Sinne darstellten, waren ebenfalls in dieser Kategorie zu finden. Unter prominenten Strukturen konnte man beispielsweise ein unspezifisch verbreitertes Mediastinum oder Lungenhili, eine verstärke Aortenelongation oder Aortensklerose, als auch periepikardiale Schwielen, verdichtetes Lungenparenchym oder vermehrte Kalkablagerung in den Gefäßen verstehen. Demnach sollten alle Veränderungen der Anatomie oder Anomalien des Thoraxröntgenbildes, die am ehesten einem altersentsprechenden und/oder physiologischen Befund des Patienten entsprachen, in das Training des Algorithmus des maschinellen Lernens miteinbezogen werden. Diese häufig auftretenden Befunde im Röntgenbild sollte der Algorithmus hinsichtlich seiner Detektionsrate bezogen auf Lungenrundherde als irrelevant identifizierten können und diese somit seine Leistung nicht negativ beeinflussen.

Eine Erweiterung der Kategorie Unauffällig stellte die Kategorie Unauffällig + Mamillenschatten dar. Diese Kategorie enthielt dieselben formalen Rahmenbedingungen, wie die zuvor erläuterte Kategorie Unauffällig. Diese Kategorie wurde jedoch zusätzlich dadurch ausgezeichnet, dass die in ihr enthaltenen Thoraxröntgenbilder sehr prominente Mamillenschatten aufwiesen. Den Nachweis, dass es sich bei der rundlichen Formation und der typischen Lokalisation der Mamillenschatten im Thoraxröntgenbild um keine pathologischen Auffälligkeiten handelte, wurde anhand vorliegender CT-Aufnahmen verifiziert.

Die Kategorie Unauffällig + beinhaltete Thoraxröntgenbilder von Patienten, die keine Lungenrundherde aufwiesen, aber eine andere Erkrankung in der Aufnahme des

Thorax sichtbar war. ‚+' bedeutete somit das Vorhandensein einer weiteren Erkrankung, die sich im Thoraxröntgenbild darstellte. Zu diesen Erkrankungen gehörten unter anderem das Vorliegen von Pleuraerguss, Pneumonie, Rippenfraktur, Hemithorax, Pneumothorax, Hämatothorax, und Emphysem. Diese Erkrankungen könnten die Leistung des Algorithmus einschränken und somit die Detektion der Lungenrundherde im Thoraxröntgenbild erschweren. Damit eine differenzierte Betrachtung dieser Thoraxröntgenbilder möglich war, enthielt diese Unterkategorie alle diesen Kriterien entsprechenden Thoraxröntgenbilder.

Eine weitere Kategorie, in der sich Thoraxröntgenbilder von Patienten ohne Lungenrundherde befanden, ist die Kategorie *Unauffällig + Fremdmaterial.* In dieser Unterkategorie waren Thoraxröntgenbilder, die keine Lungenrundherde, jedoch aber Fremdmaterial, als auch ggf. weitere Erkrankungen enthielten. Der Zusatz *FM* bedeutete Fremdmaterial. Das Fremdmaterial konnte sich im oder auf dem Thorax befinden. Häufig lag Fremdmaterial in Form von EKG-Elektroden, Clips, Portsysteme, Herzschrittmacher oder aber Schmuck des Patienten im Röntgenbild vor.

In der zweiten Oberkategorie, die als Rundherd bezeichnet wurde, waren ausschließlich Thoraxröntgenbilder mit Lungenrundherden enthalten. Analog zu der Oberkategorie Unauffällig, ließ sich diese Kategorie in Rundherd, Rundherd + und Rundherd + Fremdmaterial unterteilen.

Eine Abweichung der Unterteilung beider Oberkategorien bestand lediglich darin, dass in der Oberkategorie Rundherd keine separate Unterkategorie von Thoraxröntgenbildern mit prominenten Mamillenschatten enthalten war. Diese Ausnahme ließ sich durch die Schwierigkeit der klaren Abgrenzung zwischen Lungenrundherd und Mamillenschatten erklären. Bei Thoraxröntgenbildern, die Lungenrundherde und zugleich den Verdacht auf prominente Mamillenschatten aufzeigten, bedurfte es zumeist einer weiteren diagnostischen Abklärung des Mamillenschattens. Somit war eine eindeutige Zuordnung der Patienten nur erschwert möglich und konnte bei der Kategorisierung in Unterkategorien nicht berücksichtigt werden.

Tabelle 8: Übersicht der Über- und Unterkategorien von *Unauffällig* und *Rundherd*

Oberkategorie Unauffällig		**Oberkategorie Rundherd**	
Unterkategorie	**Abk.**	**Unterkategorie**	**Abk.**
Unauffällig	U	Rundherd	RH
Unauffällig + Mamillenschatten	U + M	Rundherd +	RH +
Unauffällig +	U +	Rundherd + Fremdmaterial	RH + FM
Unauffällig + Fremdmaterial	U + FM	-	-

Der finale Bilddatensatz mit der numerische Zusammensetzung der einzelnen Unterkategorien wird in Kapitel 5.1 erläutert.

4.3.1 Verwendung der Daten

Im nachfolgenden Kapitel wird auf die Weiterverarbeitung der Daten, deren Verwendung als Trainingsdaten, als auch die Zusammensetzung des Datensatzes für die Reader Studie näher eingegangen. Darauf folgt die Erläuterung über das Vorgehen zur Analyse der diagnostischen Genauigkeit im Rahmen der Reader Studie und unter dem Einfluss von Fremdmaterial zur Bewertung der Praxistauglichkeit.

Erläuterung über die Aufteilung des Datensatzes zum Training und zur Reader Studie

Der Datensatz der mittels Stichwortsuche und den dazugehörigen Thoraxröntgenbildern des PACS, wie in Kapitel 4.2 erläutert, erstellt wurde, besteht aus 2470 Thoraxröntgenbildern (vgl. Tabelle 14). Im Folgenden wird die Aufteilung und Verwendung dieses Datensatzes näher betrachtet. Allgemein gilt, dass die in Tabelle 8 genannten Unterkategorien aus diesem hervorgehen und dessen numerische Zusammensetzung in Kapitel 5.1 dargestellt ist.

Unter Berücksichtigung der zum Training, Validierung und Reader Studie erforderlichen numerischen Zusammensetzung von Thoraxröntgenbildern mit Lungenrundherden und Thoraxröntgenbildern ohne Lungenrundherde wurden zufällig 367 Thoraxröntgenbilder aus dem Datensatzes ausgewählt. Die genaue Verwendung wird im nachfolgenden Abschnitt detailliert erläutert.

Die 367 Thoraxröntgenbilder wurden durch 154 Thoraxröntgenbilder der japanischen Datenbank JSRT ergänzt (Shiraishi et al., 2000a). Somit bildeten insgesamt 521 Thoraxröntgenbilder die Gesamtheit der Daten, die im Anwendungsbereich des Trainings, der Validierung und der Reader Studie verwendet wurden. Diese Zusammensetzung ist im Folgenden als RetinaNet-Datensatz bezeichnet. Zunächst wurden die einzelnen Datensätze, die zum späteren Zeitpunkt für die Reader Studie und die Validierung verwendet werden sollten, aus dem RetinaNet-Datensatz entnommen. 75 Thoraxröntgenbilder bildeten den Datensatz für die Reader Studie und 18 Thoraxröntgenbilder wurden dem Validierungs-Datensatz zugesprochen. Die numerische Zusammensetzung der Datensätze wurde unter Berücksichtigung der einzelnen Unterkategorien zufällig auserwählt.

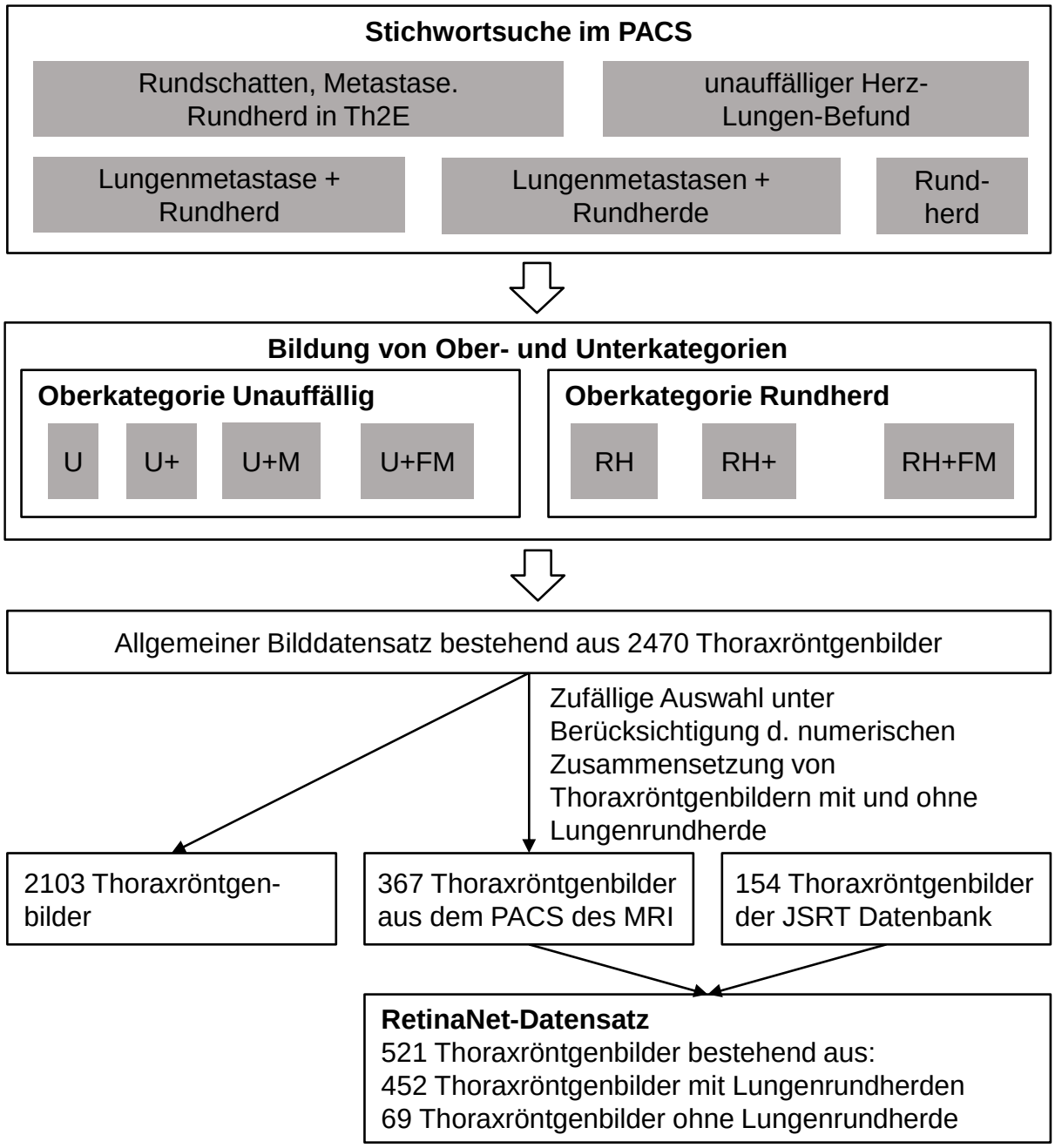


Abbildung 12: Abbildung zur Erstellung des RetinaNet-Datensatzes

Der RetinaNet-Datensatz mit 521 Thoraxröntgenbildern enthielt 452 Thoraxröntgenbildern, die mindestens einen Lungenrundherd aufwiesen. 69 Thoraxröntgenbilder wurden als Kontrolle, definiert durch das Nicht-Vorhandensein von Lungenrundherden, vorgesehen. 39 der 69 Thoraxröntgenbilder ohne Lungenrundherd waren in dem Datensatz der Reader Studie enthalten. Die verbliebenen 30 unauffälligen Thoraxröntgenbilder fanden sich im Trainingsdatensatz wieder.

Vor Durchführung der Reader Studie wurden insgesamt 428 Bilder zum Training verwendet. Hierbei wurden 30 Thoraxröntgenbilder ohne Lungenrundherd und 398 Thoraxröntgenbilder mit Lungenrundherd eingeschlossen. 154 Thoraxröntgenbilder mit Lungenrundherden stammten aus einer externen Datenbank (JSRT) (Shiraishi et al.,

2000a). Die verbliebenen 244 Thoraxröntgenbilder mit Lungenrundherden wurden aus dem PACS des Klinikums rechts der Isar extrahiert, und die Lungenrundherde der Thoraxröntgenbilder im RetinaNet-Datensatz wurden annotiert. Von 521 Thoraxröntgenbildern waren 154 Thoraxröntgenbilder der JSRT Datenbank bereits mit Informationen zur Detektion von Lungenrundherden versehen, sodass eine Annotierung dieser nicht notwendig war (Shiraishi et al., 2000a). Wie bereits vorab erläutert, enthielten von den verbliebenen 367 Thoraxröntgenbildern des RetinaNet-Datensatzes 69 Thoraxröntgenbilder keinen Lungenrundherd. Somit wurden bei insgesamt 298 Thoraxröntgenbilder suspekte Lungenrundherde händisch auf Pixelebene durch einen Radiologen mit drei Jahren Berufserfahrung detailliert markiert. 244 der Thoraxröntgenbilder wurden zum Training und 18 Thoraxröntgenbilder für die Validierung des RetinaNet verwendet. 36 der 244 Thoraxröntgenbilder waren Bestandteil der Reader Studie. Der Radiologe, der für die Annotierung der Thoraxröntgenbilder verantwortlich war, wurde nicht in die Reader Studie eingeschlossen. Zum Zeitpunkt der Annotierung waren die Bilder bereits anonymisiert und von weiteren Patientendaten isoliert. Somit wurden diese ohne das Vorliegen von Befunden oder Patienteninformationen, wie Alter, Geschlecht und Vorerkrankung untersucht und, die als suspekt zu wertenden Lungenrundherde, identifiziert. Durch ein hauseigenes Programm war es möglich, bei der anonymen Befundung der Thoraxröntgenbilder einen beliebigen Zoom einzustellen und somit Lungenrundherde verschiedener Größe und Formen auf den Pixel genau zu markieren. Die händisch ausgeführte rand- und pixelgenaue Markierung der suspekten Regionen diente dazu, diese mit der Funktion des Algorithmus des maschinellen Lernens zu vergleichen. Um die Testergebnisse des trainierten Modells zu überprüfen, konnte ein weiteres System die Bereiche, in denen der Algorithmus einen Rundherd erkannt hat, visuell darstellen. Hierbei wurde ein Rechteck um den erkannten Bereich gezeichnet. Zudem wurde ein Score, welcher angab, wie sicher sich der Algorithmus bei der Detektion von Lungenrundherden verhielt, errechnet. Anhand der vorab erfolgten händischen Annotierung der Lungenrundherde konnte ein direkter Vergleich zwischen der Markierung von Lungenrundherden durch den Algorithmus und die Detektion von Lungenrundherden von ärztlicher Seite aus entstehen.

Zusammensetzung der Daten für die Reader Studie

Nachdem im vorherigen Abschnitt die für den weiteren Ablauf der Methodik bedeutenden Schritte erläutert worden sind, wird in dem nachfolgenden Abschnitt auf die finale Zusammensetzung der Daten für die Reader Studie eingegangen.

Tabelle 9 zeigt die Zusammensetzung des Datensatzes der Reader Studie. Insgesamt wurden in der Reader Studie 75 Thoraxröntgenbilder von 75 unterschiedlichen Patienten eingeschlossen. Dies impliziert, dass die Lungenrundherdgröße und deren Lokalisation innerhalb des Thoraxröntgenbildes inter- und intraindividuell sehr variabel waren. 36 Thoraxröntgenbilder gehörten zu der Oberkategorie Rundherd und insgesamt

39 Thoraxröntgenbilder repräsentierten die Oberkategorie Unauffällig. Die Auswahl der Bilder innerhalb der jeweiligen Unterkategorien erfolgte zufällig unter Berücksichtigung der vorab definierten numerischen Zusammensetzung. Die Unterkategorie Rundherd, sowie die Unterkategorie Unauffällig stellten mit je 20 Thoraxröntgenbildern mehr als 50 % der Daten für die Reader Studie dar. Auf den 20 Thoraxröntgenbildern der Unterkategorie Rundherd ließen sich insgesamt 33 Lungenrundherde erkennen. Im Schnitt entsprach das 1,65 Lungenrundherde pro Thoraxröntgenbild. In der zufälligen Auswahl der 75 Bilder für die Reader Studie wurden keine Thoraxröntgenbilder mit zusätzlichen Erkrankungen (RH + und U +) ausgewählt und somit wurden diese in der Reader Studie nicht miteingeschlossen. Es wurden jeweils 16 Thoraxröntgenbilder für die Unterkategorien mit Fremdmaterial herangezogen. Die 16 Thoraxröntgenbilder der Unterkategorie Rundherd + Fremdmaterial zeigten insgesamt 29 Lungenrundherde. Dies entsprach einem Schnitt von 1,8 Lungenrundherde pro Thoraxröntgenbild mit Fremdmaterial.
Die Unterkategorie Unauffällig + Mamillenschatten setzte sich aus 3 Thoraxröntgenbildern zusammen, die mindestens einen Mamillenschatten pro Aufnahme enthielten. Aufgrund der Ähnlichkeit von Mamillenschatten und Lungenrundherd stellte dieses Vorkommnis somit eine besondere Herausforderung für die korrekte Befundung dar. Sowohl die Unterkategorie U +, als auch die Unterkategorie RH + wurden nicht in die Reader Studie mit eingeschlossen.

Tabelle 9: Zusammensetzung der Thoraxröntgenbilder für die Reader Studie

	RH	RH +	RH +FM	U	U +	U + FM	U + M
Anzahl d. Bilder	20	0	16	20	0	16	3
Anzahl d. Lungenrundherde	33	0	29	-	-	-	-

Zusammensetzung der Daten für die Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial

Basierend auf den in SCHOBER's Arbeit annotierten Daten wurde in Kapitel 5.5 dieser Arbeit weiter untersucht, inwieweit das ResNet zur Klassifikation von Bildern durch Fremdmaterial beeinflusst wurde. Tabelle 10 zeigt die Zusammensetzung des Datensatzes, der zur Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial verwendet wurde. Dieser setzte sich aus insgesamt 55 Rundherd positiven Bildern und 57 unauffälligen Thoraxröntgenbilder zusammen, wobei 41 der Thoraxröntgenbilder Fremdmaterial enthielten. 17 Thoraxröntgenbilder entstammten aus der Kategorie Rundherd + Fremdmaterial. 24 Thoraxröntgenbilder, als gesunde Kontrolle, gehen aus der Unterkategorie Unauffällig + Fremdmaterial hervor.

Tabelle 10: Datensatz zur Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial

Gruppe	RH	RH +	RH +FM	U	U +	U + FM	Gesamt
Anzahl d. Bilder	29	9	17	32	1	24	112

Da es sich bei diesem Datensatz um eine Screening-Untersuchung handelt, wurden die Thoraxröntgenbilder durch einen Radiologen mit 3 Jahren Berufserfahrung auf das Vorkommen von Lungenrundherden überprüft, jedoch wurde auf eine pixelgenaue Annotierung verzichtet, da die verwendete Architektur (ResNet) keine pixelgenaue Ausgabe erforderte (Schultheiss et al., 2020), (Schober, 2019) Tabelle 11 zeigt die Anzahl der Thoraxröntgenbilder mit Fremdmaterial, die aus dem zuvor genannten Datensatz entnommen wurden und weiterführend in SCHOBER ET AL. für das RetinaNet untersucht wurden.

Tabelle 11: C1 Datensatz zur Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial in Anlehnung an (Schober, 2019)

	Anzahl der Bilder mit Fremdmaterial aus Screening Datensatz C1
Anzahl Thoraxröntgenbilder mit LRH (RH + FM)	17
Anzahl Thoraxröntgenbilder ohne LRH (U + FM)	24
Gesamt	41

Die 41 Thoraxröntgenbilder, die Fremdmaterial aufwiesen, wurden erneut untersucht und die verschiedenen Fremdmaterialien manuell identifiziert. Anhand des gehäuften Vorkommens der gleichen Art von Fremdmaterial wurde dieser Datensatz in drei verschiedene Fremdmaterial-Kategorien unterteilt.
Die erste Kategorie besteht aus 11 Thoraxröntgenbildern auf denen EKG-Elektroden sichtbar waren. Eine zweite Kategorie mit insgesamt 14 Thoraxröntgenbildern war gekennzeichnet durch das Vorhandensein eines Portsystems. Die größte Kategorie mit 35 Bildern bildeten Thoraxröntgenbilder in denen aufliegendes Fremdmaterial des Patienten zu erkennen war. Hierzu gehörten beispielweise Kleidungsstücke, wie Knöpfe und Unterwäsche oder Schmuck, wie Piercings, Halsketten oder Ohrringe.
Tabelle 12 zeigt die numerische Zusammensetzung der zuvor genannten drei Kategorien. Zu berücksichtigen galt es, dass unterschiedliche Fremdmaterialien innerhalb eines Thoraxröntgenbilder vorhanden sein konnten. Somit waren insgesamt 61

Fremdmaterialien auf insgesamt 41 Thoraxröntgenbilder zu finden. Im Schnitt bedeutete das 1,49 Fremdmaterialien pro Thoraxröntgenbild.

Tabelle 12: Anzahl von Thoraxröntgenbildern mit unterschiedlichem Fremdmaterial

	Anzahl von Thoraxröntgenbildern mit FM		
	EKG-Elektroden	**Portsystem**	**Aufliegendes Fremdmaterial**
Anzahl von Thoraxröntgenbildern mit LRH	1	9	12
Anzahl von Thoraxröntgenbildern ohne LRH	11	5	23
Gesamt Anzahl von vorhandenem Fremdmaterial	12	14	35

4.3.2 Vorgehen zur Analyse der diagnostischen Genauigkeit zur Detektion von Lungenrundherden im Rahmen der Reader Studie

Um den klinischen Einsatz von Algorithmen, die Lungenrundherde in Thoraxröntgenbilder identifizieren, zu bewerten, wurde eine Reader Studie durchgeführt. Im folgenden Abschnitt sind hierzu die Rahmenbedingungen erläutert und die Funktionsweise des Algorithmus in einem realistischen medizinischen Setup geprüft. Die Darlegung der Ergebnisse der Reader Studie erfolgt in Kapitel 5.3 und wird in Kapitel 6.2 diskutiert.

Die Reader Studie diente dem Vergleich der Detektion von Lungenrundherden durch einen Algorithmus des maschinellen Lernens und zweier Radiologen. Hiermit wurde analysiert, ob der Algorithmus zur Detektion von Lungenrundherden in der Praxis geeignet ist.

Zwei Radiologen mit vier und sechs Jahren Berufserfahrung analysierten mittels eines hauseigenen Annotationsprogramms den Datensatz der Reader Studie, der in Kapitel 4.3.1 aufgeführt wurde. Das Annotationsprogramm ermöglichte es einzelne Pixel auf einem radiologischen Bild zu markieren und diese Markierung zu speichern. Somit konnte die händische Markierung mit der Markierung des Algorithmus verglichen werden. Die Markierung eines Lungenrundherds seitens der Radiologen galt als positiv, wenn der Radiologe mindestens einen Pixel des Lungenrundherds markiert hatte. Der Annotationsprozess für diese Studie bezog sich lediglich auf die Markierung von Lungenrundherden, ohne Beachtung weiterer Auffälligkeiten der Thoraxröntgenbilder. Diese Markierungen wurden auf einer Karte, die die gleichen Abmessungen wie das originale Thoraxröntgenbild hatte, gespeichert. Im Anschluss konnte diese Karte mit

den Markierungen des Algorithmus verglichen werden. Die Annotierung und Detektion von Lungenrundherden seitens der Radiologen sollten die täglichen Bedingungen der Befundung in der Radiologie simulieren. Aus diesem Grund arbeiteten die Radiologen mit einer Zeitrestriktion von 10 Sekunden pro Bild, um alle Lungenrundherde eines Thoraxröntgenbildes zu markieren. 10 Sekunden entspricht der durchschnittlichen Zeit, die ein geübter Radiologe für die Entscheidung einer weiterführenden Diagnostik auf Grundlage der Radiographie benötigt, um die Malignität durch eine erweiterte Bildgebung, bspw. im Rahmen einer CT-Untersuchung einzugrenzen.

4.3.3 Vorgehen zur Analyse der diagnostischen Genauigkeit zur Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial

Um mögliche Schwierigkeiten des Algorithmus des maschinellen Lernens bezogen auf die Detektion von Lungenrundherden bei gleichzeitigem Vorkommen von Fremdmaterialien im Thoraxröntgenbild zu ermitteln, wird in dem folgenden Abschnitt diese Situation genauer untersucht.

Wie bereits vorab erläutert, wurde die Annotierung der Thoraxröntgenbilder, die Lungenrundherde enthalten, durch einen Radiologen mit drei Jahren Berufserfahrung im Rahmen der RetinaNet Verwendung durchgeführt. Anschließend wurden die durch den Algorithmus identifizierten Lungenrundherde, mit den Annotierungen des Radiologen, auf ihre Richtigkeit hin, überprüft und die Leistung verglichen. Durch die Kategorisierung der Thoraxröntgenbilder in Abhängigkeit vom Vorkommen von verschiedenen Fremdmaterialen konnte hierbei spezifisch untersucht werden, ob besondere Fremdmaterialien potentielle Störfaktoren für die Leistung des Algorithmus darstellten. Die Ergebnisse dieser Untersuchung sind in Kapitel 5.5 dargestellt und werden in Kapitel 6.2 diskutiert.

4.4 Ergebnisdarstellung

Im folgenden Kapitel wird die Herangehensweise zur Evaluation der Ergebnisse erläutert. Hierbei werden die Grundparameter der Vierfeldertafel, als auch der F2 Score und die Darstellung der Ergebnisse mittels ROC-Kurve beschrieben.

4.4.1 Herangehensweise zur Evaluation der Ergebnisse

Um die Leistung eines maschinellen Lernmodells zu bewerten, gibt es mehrere Indikatoren, die die Güte eines Klassifikators quantifizieren können.

Wie bereits in Kapitel 2.5.1 erläutert, zählen TP, FP, FN und TN zu den quantitativen Grundparametern der Vierfeldertafel. Im Folgenden werden diese auf die Evaluation der Ergebnisse angewendet und dienen der Auswertung der Funktion des Algorithmus. Sie stellen bezogen auf die Detektion der Lungenrundherde im Thoraxröntgenbild eine Möglichkeit dar, die Übereinstimmung der Markierung durch den Algorithmus, mit dem tatsächlichen Vorkommen von tumorösen Gewebe zu beschreiben.

Im Rahmen dieser Arbeit entspricht die Bezeichnung ‚Krankheit liegt vor' dem Vorliegen von Lungenrundherden im Thoraxröntgenbild. Die Aussage, ob der Test positiv oder negativ ist, ist hierbei mit der Leistung des Algorithmus in Hinsicht auf die Detektion von Lungenrundherden gleichzusetzen. Hierbei beziehen sich die Ergebnisse auf die Detektion eines einzelnen Lungenrundherds und nicht auf das gesamte Thoraxröntgenbild.
Ein TP Ergebnis bedeutet, dass ein Lungenrundherd vorliegt und der Algorithmus diesen als solchen identifiziert. Ein FP Ergebnis hingegen sagt aus, dass der Algorithmus fälschlicherweise Lungenrundherde detektiert, obwohl auf dem Thoraxröntgenbild diese nicht vorliegen. Die Summe von TP und FP schließt somit alle Testpositiven ein, unabhängig vom Vorliegen von Lungenrundherden. Hinsichtlich der Detektion von Lungenrundherden bedeutet FN, dass der Algorithmus keine Lungenrundherde detektiert, obwohl diese im Thoraxröntgenbild vorliegen und detektiert werden hätten müssen. TN bedeutet hingegen, dass keine Lungenrundherde vorliegen und der Algorithmus ebenso keine detektiert, also ein richtiges Ergebnis erbringt. TN beschreibt demnach nicht-tumoröses Gewebe, dass vom Algorithmus korrekterweise als nicht-tumorös klassifiziert wurde und ist somit eine Beschreibung, die in diesem Kontext keine Relevanz aufweist, da dieses nicht eindeutig messbar ist.
Die Summe von FN und TN beschreibt alle testnegativen Ergebnisse. Testnegativ bedeutet in dieser Arbeit, dass der Algorithmus keine Lungenrundherde in den zu bewertenden Thoraxröntgenbildern detektiert.
TP und FN stellen gemeinsam die Gruppe aller Erkrankten dar. In dieser Arbeit entspricht das allen Thoraxröntgenbildern, die Lungenrundherde enthalten. FP und TN hingegen bilden die Gesamtheit der Thoraxröntgenbilder, die keine Lungenrundherde enthalten. Die Summe von TP, FP, TN und FN bildet die Gesamtheit aller berücksichtigten Thoraxröntgenbilder.
Das Maß der Genauigkeit beschreibt in dieser Arbeit die Anzahl der korrekten Klassifizierungen durch den Algorithmus, bezogen auf die Anzahl aller untersuchten Thoraxröntgenbilder.
Die TPR (Sensitivität, Recall) gibt an mit welchem Prozentsatz das Vorliegen von Lungenrundherden auch tatsächlich durch den Algorithmus erkannt wird.
Die FPR beschreibt den Anteil der fälschlicherweise positiv auf Lungenrundherde analysierten Thoraxröntgenbilder durch den Algorithmus, obwohl diese unauffällig sind.
Die FNR beschreibt hingegen den Anteil der Thoraxröntgenbilder, die tatsächlich Lungenrundherde aufweisen, diese jedoch vom Algorithmus nicht identifiziert und somit als unauffällig deklariert wurden.
Die TNR bzw. Spezifität gibt die Wahrscheinlichkeit an, dass unauffällige Thoraxröntgenbilder, also ohne das Vorliegen von Lungenrundherden, auch als unauffällig deklariert werden. Anders ausgedrückt, stellt sie die Wahrscheinlichkeit dar, dass tatsächlich gesunde Patienten, die keine Lungenrundherde aufweisen, durch den Algorithmus

des maschinellen Lernens ebenso als gesund erkannt werden. Wie bereits vorab erläutert ist die Relevanz der TNR bzw. Spezifität in dieser Arbeit von untergeordneter Bedeutung und wird hier nicht weiter berücksichtigt. Die TNR kann nicht allumfassend gesondert berechnet werden, sondern ergibt sich im Ausschlussverfahren durch die anderen Parameter. Diese Rate beschreibt alle verbliebenen möglichen Begrenzungsrahmen, die das unauffällige Umfeld innerhalb eines Thoraxröntgenbildes beschreiben. Hieraus ergeben sich Millionen von Möglichkeiten.
Anhand der Anzahl von TP, FP, und FN Klassifikationen wird die diagnostische Genauigkeit des Algorithmus und zweier Radiologen in der Reader Studie bestimmt und miteinander verglichen.

4.4.1 ROC-Kurve

Um den Gesundheitszustand und in diesem Fall das Vorkommen von Lungenrundherden des Patienten angemessen vorhersagen zu können, sollte ein idealer Algorithmus des maschinellen Lernens eine hohe TPR und zugleich eine niedrige FPR erreichen. Diese Beziehung zwischen beiden Eigenschaften wird in Kapitel 2.5.3 mittels einer ROC-Kurve dargestellt und durch die AUC quantifiziert.
Die Güte der Algorithmen des maschinellen Lernens werden dadurch bestimmt, wie nah die jeweiligen ROC-Kurven mit TPR = 1 und FPR = 0 an der oberen linken Ecke lokalisiert sind. Liegen die ROC-Kurven dieser sehr nah, so kann man in Hinsicht auf dieses Arbeit die Aussage treffen, dass der Algorithmus alle Patienten mit Lungenrundherden und alle Patienten ohne Lungenrundherde korrekt ermittelt hat. Um diese Eigenschaft zu quantifizieren, wird die AUC berechnet. Nimmt auch diese Werte von AUC = 1 an, kann der Algorithmus den Gesundheitszustand bezogen auf Lungenrundherde genau bestimmen. Die ROC-Kurve veranschaulicht in dieser Arbeit den Vergleich der Detektion von Lungenrundherden unter dem Einfluss von verschiedene Fremdmaterialien. Anhand dieser Ergebnisdarstellung soll übersichtlich verglichen werden, bei welchem Vorliegen von Fremdmaterial der Algorithmus Schwierigkeiten aufweist, Lungenrundherde verlässlich zu detektieren. Mittels Wertepaaren von (TPR/FPR) wird für jede Kategorie von Fremdmaterial (EKG-Elektroden, Portsystem, aufliegendes Fremdmaterial), als auch für alle Patienten des untersuchten Datensatzes jeweils eine ROC-Kurve in einem gemeinsamen Koordinatensystem generiert, sodass ein unmittelbarer Vergleich dieser Kategorien stattfinden kann. Aus der ROC-Kurve lässt sich daraufhin die AUC für jede Kategorie berechnen. Tabelle 13 zeigt das Konzept der Konfusionsmatrix. Die Tumorvorhersage (engl. Prediction) entspricht der Vorhersage durch den Algorithmus und die Grundwahrheit (engl. Groundtruth) entspricht der wahren Gegebenheit.

Tabelle 13: Konzept der Konfusionsmatrix für die Tumorvorhersage

		Groundtruth	
		Tumor	**No Tumor**
Prediction	**Tumor**	True Positive TP	False Positive FP
	No Tumor	False Negative FN	True Negativ TN

4.4.2 F2 Score

Der F2 Score ist eine Kennzahl, die verschiedene Parameter zur Evaluation der Funktion von Algorithmen und Radiologen beinhaltet und somit mittels einer einzigen Kennzahl einen Überblick über die Leistung dieser geben kann.

Wie bereits erläutert setzt sich der F2 Score aus der Kombination von Trefferquote bzw. TPR und Genauigkeit mithilfe des gewichteten harmonischen Mittels zusammen. Speziell im F2 Score wird die Trefferquote höher als die Genauigkeit gewichtet. (Minto, 2009) Diese Tatsache ist in medizinischen Anwendungsbereichen von Vorteil, da die Detektion jedes Lungenrundherdes (positive Beobachtung) von Bedeutung ist und diese die Priorität bei der Diagnostik darstellt. Somit wird eine Messgröße erstellt, die Genauigkeit und TPR kombiniert, diese jedoch unterschiedlich gewichtet.

Der F2 Score gibt bei der Detektion von Lungenrundherden eine Kennzahl an, die den Vergleich von Untersuchungen übersichtlich ermöglicht. Jedoch zeigt der F2 Score eine Invarianz. Verändert sich beispielsweise die TPR ins Positive und die Genauigkeit ins Negative, so kann trotzdem der gleiche F2 Score bestehen bleiben und es kann keine direkte Aussage über die Detektionsrate gemacht werden. Zudem kann der F2 Score ungenau sein, wenn ein extremes Ungleichgewicht zwischen tatsächlich positiven und negativen Fällen vorliegt, wie bei dem Vorkommen von seltenen Krankheiten. Ebenso kann die Kenngröße verfälscht werden, wenn die Anzahl der an einem Test teilnehmenden Kranken deutlich geringer, als die der Gesunden ist und es somit nur zu einem gering positiven Vorhersagewert kommen kann.

Dennoch ist der F2 Score ein geeignetes Mittel, um die allgemeine Funktion von Algorithmen und Radiologen im Bereich des maschinellen Lernens zu vergleichen und wird in dieser Dissertation zur Ergebnisdarstellung verwendet.

5 Ergebnisse

Das fünfte Kapitel zeigt zunächst die numerische Zusammensetzung der einzelnen Unterkategorien in Tabelle 14. Anschließend werden jeweils die Ergebnisse der Detektion von Lungenrundherden in der radiologischen Befundung als auch des maschinellen Lernens vorgestellt. Darauf aufbauend folgt der Vergleich der Detektionsleistung der Radiologen mit dem Algorithmus des maschinellen Lernens. Abschließend erfolgt die Evaluation der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial mit der auf maschinellem Lernen basierten Methode.

5.1 Numerische Zusammensetzung der Unterkategorien

Einteilung des Datensatzes in Unterkategorien

Im folgenden Abschnitt wird die Anzahl der Thoraxröntgenbilder in den einzelnen Unterkategorien tabellarisch dargestellt. Die formalen Bedingungen dieser, wurden bereits in Kapitel 4.3 aufgeführt. Die Zusammensetzung der einzelnen Kategorien und die Verwendung dieser Daten galt als Basis für das Training und die Evaluation der diagnostischen Genauigkeit des Algorithmus des maschinellen Lernens. Alle hier berücksichtigten Thoraxröntgenbilder zeigten sich in der Darstellung posterior-anterior.

Tabelle 14: Anzahl der Thoraxröntgenbilder und Zeitraum der Aufnahmen der Unterkategorien

	Anzahl d. Thorax-röntgenbilder	**Zeitraum d. Aufnahmen**
Unauffällig	1790	01.01.2015 – 03.03.2018
Unauffällig +	5	14.04.2015 – 26.08.2016
Unauffällig + FM	148	27.10.2010 – 28.02.2018
Unauffällig + Mamillenschatten	72	07.01.2010 – 10.11.2017
Gesamt unauffällig	2015	07.01.2010 – 03.03. 2018

	Anzahl d. Thorax-röntgenbilder	**Zeitraum d. Aufnahmen**
Rundherd	195	18.02.2004 – 31.01.2018
Rundherd +	57	12.02.2004 – 22.12.2017
Rundherd + Fremdmaterial	203	13.01.2004 – 26.03.2018
Gesamt mit Rundherd	455	13.01.2004 – 26.03.2018

Tabelle 14 stellt die beiden Oberkategorien Unauffällig und Rundherd mit ihren jeweiligen Unterkategorien dar. Die tabellarische Form gibt einen Überblick über die Anzahl der Thoraxröntgenbilder, sowie den Zeitraum der Aufnahmen. Die Summe der Thoraxröntgenbilder aller Kategorien entspricht einer Anzahl von 2470 und die Verwendung dieser wurde bereits in Kapitel 4.3.1 näher erläutert.

5.2 Ergebnisse der Detektion von Lungenrundherden der radiologischen Befundung

Die Datenzusammensetzung der Reader Studie, die bereits in Kapitel 4.3.1 in Tabelle 9, dargestellt wurde, bezieht sich auf die Ergebnisse der Kapitel 5.2, 5.3 und 5.4 und ist übersichtshalber im Folgenden noch einmal als Tabelle 15 dargestellt.

Tabelle 15: Zusammenstellung der Thoraxröntgenbilder für die Reader Studie

	RH	RH +	RH +FM	U	U +	U + FM	U + M
Anzahl d. Bilder	20	0	16	20	0	16	3
Anzahl d. Lungenrundherde	33	0	29	-	-	-	-

Die in Tabelle 16 dargestellten Ergebnisse der Reader Studie zeigen eine Übersicht über die Detektion von Lungenrundherden in der radiologischen Befundung. Hierbei werden die Anzahl für TP, FP und FN Ergebnisse, als auch der F2 Score für Radiologen A und B angegeben. Insgesamt galt es 62 Lungenrundherde zu detektieren.
Radiologe A detektierte 43 Lungenrundherde, das entsprach einer Detektionsrate von 69,4 %. Insgesamt übersah der Radiologe somit 19 der 62 zu detektierenden Lungenrundherde. 29 Regionen identifizierte er fälschlicherweise als Lungenrundherde. Diese Daten berücksichtigt, ergab sich für Radiologe A ein F2 Score von 0,672.
Radiologe B identifizierte 39 von 62 Lungenrundherden. Dies entsprach einer Detektionsrate von 62,9 %. 23 von 62 Lungenrundherden wurden von ihm übersehen. Zudem bezeichnete der Radiologe B 30 Regionen fälschlicherweise als Lungenrundherde. Der F2 Score betrug hier 0,615.
Anhand der F2 Scores ließ sich feststellen, dass die Leistung beider Radiologen innerhalb der Reader Studie vergleichbar war. Radiologe A zeigte im Vergleich zu Radiologe B eine sensitivere Detektion und einen F2 Score, der um 0,057 höher war. Zudem detektierte Radiologe A 4 Lungenrundherde mehr als Radiologe B und übersah zugleich ebenso 4 Lungenrundherde weniger als Radiologe B. Die niedrigere Anzahl an FP Detektionen unterschied sich von Radiologe A zu Radiologe B nur um einen Lungenrundherd.

Tabelle 16: Ergebnisse der Detektion von Lungenrundherden in der radiologischen Befundung

Radiologe	TP	FP	FN	F2 Score
A	43	29	19	0,672
B	39	30	23	0,615

5.3 Ergebnisse der Detektion von Lungenrundherden durch den Algorithmus des maschinellen Lernens

Tabelle 17 zeigt die Ergebnisse des Algorithmus des maschinellen Lernens der Reader Studie. Die Tabelle zeigt den F2 Score, als auch die Anzahl der TP, FP und FN Ergebnisse des Algorithmus.

Tabelle 17: Ergebnisse der Detektion von Lungenrundherden durch den Algorithmus des maschinellen Lernens

	TP	FP	FN	F2 Score
RetinaNet	39	35	23	0,606

Der Algorithmus des maschinellen Lernens detektierte 62,9 % der vorhandenen Lungenrundherde. Dies entsprach einer Anzahl von 39 TP. 23 Lungenrundherde wurden jedoch durch den Algorithmus übersehen und somit nicht detektiert. Zudem identifizierte er 35 Regionen fälschlicherweise als Lungenrundherde. Auf diesen Daten basierend, ergab sich ein F2 Score von 0,606.

5.4 Vergleich der radiologischen und maschinellen Performance

Im folgenden Kapitel wird die Detektion von Lungenrundherden im Rahmen der Reader Studie verglichen. Hierbei werden die Parameter TP, FP und FN, als auch der F2 Score berücksichtigt. Im weiteren Verlauf wird dann näher auf die einzelnen Unterkategorien und die Detektionsleistung durch Radiologen und den Algorithmus des maschinellen Lernens innerhalb dieser Kategorien eingegangen. Hierbei wird anhand der TP, FP und FN Detektionen näher auf die Leistung der Radiologen A und B und des Algorithmus des maschinellen Lernens eingegangen.

Tabelle 18: Tabellarische Abbildung der Ergebnisse der Reader Studie in Anlehnung an (Schober, 2019)

Radiologe	**TP**	**FP**	**FN**	**F2 Score**
A	43	29	19	0,672
B	39	30	23	0,615
Algorithmus	**TP**	**FP**	**FN**	**F2 Score**
RetinaNet	39	35	23	0,606

Radiologe A detektierte 69,4 % aller vorhandenen Lungenrundherde innerhalb der Reader Studie und zeigte somit die beste Leistung hinsichtlich TP Detektionen. Radiologe B und der Algorithmus detektierten beide 62,9 % der Lungenrundherde der Reader Studie, wiesen jedoch unterschiedliche F2 Scores auf. Tabelle 18 zeigt, dass verglichen zwischen Radiologe A und B, und dem Algorithmus, Radiologe A den größten F2 Score, als auch die größte Sensitivität im Rahmen der Reader Studie erzielte. Radiologe A detektierte 4 Lungenrundherde mehr als Radiologe B und der Algorithmus. Somit unterschied sich auch die Anzahl FN Ergebnisse um 4. Radiologe B und der Algorithmus übersahen 4 Lungenrundherde und hatten dadurch eine höhere Anzahl an FN Ergebnissen, als Radiologe A. Die Anzahl FP Ergebnisse unterschied sich von Radiologe A zu Radiologe B um eine Region. Diese wurde also von Radiologe B fälschlicherweise als Lungenrundherd eingestuft. Der Algorithmus hingegen klassifizierte 5 weitere Regionen als Lungenrundherd und identifizierte somit fälschlicherweise 6 Regionen mehr als Lungenrundherde, als Radiologe A. Der F2 Score von Radiologe A zum Algorithmus zeigte eine Differenz von 0,066, wohingegen sich der F2 Score von Radiologe B zum Algorithmus nur um 0,009 unterschied. Radiologe B und der Algorithmus zeigten dieselbe Anzahl an TP und FN Detektionen. Demnach kam der Unterschied zwischen den F2 Scores von Radiologe B und dem Algorithmus durch die unterschiedliche Anzahl an FP Ergebnissen zustande. Der Algorithmus identifizierte fälschlicherweise 5 Regionen mehr als Lungenrundherd, verglichen zu Radiologe B. Von den Lungenrundherden, die mindestens einer der beiden Radiologen detektierte, konnte der Algorithmus 32 Lungenrundherde identifizieren. In der Arbeit von SCHOBER konnte der Algorithmus insgesamt 7 von 15 Lungenrundherden identifizieren, die keiner der beiden Radiologen detektierte. (Schober, 2019), (Schultheiss et al., 2020)

Tabelle 19 zeigt eine detaillierte Aufschlüsselung über die Anzahl von TP, FP und FN Detektionen innerhalb der Unterkategorien der Reader Studie. Hierbei wurde die Leistung des Algorithmus des maschinellen Lernens und die der Radiologen bezogen auf die einzelnen Unterkategorien untersucht. Innerhalb der Kategorie der Rundherde

ohne weitere Besonderheiten galt es 33 Lungenrundherde zu detektieren, wohingegen die Kategorie Rundherd + Fremdmaterial 29 Lungenrundherde enthielt.
Innerhalb der Kategorie der Rundherde ohne weitere Besonderheiten zeigte Radiologe A verglichen mit Radiologe B und dem Algorithmus die größte Anzahl TP Detektionen und einen deutlichen Unterschied zur geringeren Anzahl an FP und FN Ergebnissen. Radiologe A zeigte eine Anzahl von 4 FP Detektionen, während Radiologe B 13 und der Algorithmus 10 FP Detektionen aufwiesen. Zudem übersah Radiologe A 6 Lungenrundherde, die durch die 6 FN Ergebnisse dargestellt wurden. Radiologe B zeigte 11 FN Ergebnisse und der Algorithmus wies eine Anzahl von 10 FN Detektionen auf. Der Algorithmus detektierte einen Rundherd mehr, als Radiologe B und wies zugleich 3 FP und ein FN Ergebnis weniger auf als dieser. In der Kategorie der Rundherde in Kombination mit dem Vorliegen von Fremdmaterial waren die Unterschiede zwischen den Radiologen selbst, und dem Algorithmus deutlich geringer als in der Kategorie mit ausschließlich vorliegenden Rundherden. In der Kategorie Rundherd + Fremdmaterial zeigte Radiologe B eine bessere Leistung als Radiologe A. Hierbei fiel besonders die um 4 geringere Anzahl an FP Detektionen auf. Der Algorithmus und Radiologe A zeigten dieselbe Anzahl an TP und FN Detektionen. Diese unterschieden sich um je eine Detektion der TP und FN Ergebnisse des Radiologen B. Zugleich wies der Algorithmus eine um 4 geringere Anzahl an FP Detektionen auf als Radiologe B und lieferte hiermit das beste Ergebnis bezogen auf die FPR der Unterkategorie Rundherd + Fremdmaterial. Der Algorithmus detektierte in der Unterkategorie Rundherd deutlich mehr FP Ergebnisse und in der Kategorie Unauffällig + Fremdmaterial war die FPR des Algorithmus deutlich höher, als die der Radiologen.
In der Oberkategorie der unauffälligen Thoraxröntgenbilder gab es keine zu detektierenden Lungenrundherde und somit auch entsprechend keine TP und FN Ergebnisse. Thoraxröntgenbilder der Kategorie Unauffällig, ohne weitere Besonderheiten, wurden am besten durch die Leistung des Radiologen B mit 3 FP Detektionen eingeschätzt. Radiologe A und der Algorithmus hingegen zeigten beide doppelt so viele FP Detektionen in dieser Kategorie. Dennoch waren die FP Detektionen durch den Algorithmus in der Kategorie Unauffällig um 5 Ergebnisse geringer, als in der Kategorie Unauffällig + Fremdmaterial. In der Kategorie Unauffällig + Fremdmaterial wies der Algorithmus mit 11 FP die höchste Anzahl FP Detektionen auf. Im Vergleich dazu zeigten Radiologe A und B jeweils nur 3 und 2 FP Detektionen. Vergleicht man dies mit der Oberkategorie der Rundherde, die insgesamt 3 Thoraxröntgenbilder weniger enthielt als die Kategorie Unauffällig, so zeigte sich die FPR des Algorithmus genau gegenteilig. In der Kategorie Rundherd zeigte der Algorithmus eine höhere Anzahl FP Detektionen, als in der Unterkategorie Rundherd + Fremdmaterial (RH + FM).
Die Kategorie der unauffälligen Thoraxröntgenbilder mit vorliegenden Mamillenschatten zeigten bei Radiologe A und B, als auch dem Algorithmus, mit einer Anzahl von einer FP Detektion dieselben Ergebnisse.

Tabelle 19: Tabellarische Abbildung der Kategorien-spezifischen Ergebnisse aus der Reader Studie

	RH			***RH + FM***			***U***			***U + FM***			***U + M***		
	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN
Radiologe A	27	04	06	16	15	13	-	06	-	-	03	-	-	01	-
Radiologe B	22	13	11	17	11	12	-	03	-	-	02	-	-	01	-
Retina-Net	23	10	10	16	07	13	-	06	-	-	11	-	-	01	-

5.5 Ergebnisse der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial

Im folgenden Kapitel werden die Ergebnisse der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial, wie EKG-Elektroden, implantierte Portsysteme, als auch aufliegendem Fremdmaterial dargestellt. Hierbei wurde der Datensatz, wie er in Tabelle 11 in Kapitel 4.3.1 dargestellt ist, und die Leistung des Algorithmus unter diesen Bedingungen untersucht.

Tabelle 20: Ergebnisse der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial

	Fremdmaterial	**EKG-Elektroden**	**Portsystem**	**Aufliegendes Fremdmaterial**
Anzahl FPR/TPR Koordinaten	15	2	8	11
AUC	0,836	1,000	0,756	0,815

Die Ergebnisse der unterschiedlichen Fremdmaterialien sind in Form von ROC-Kurven aufgeführt. Um diese direkt vergleichen zu können, wurden die ROC-Kurven in einem gemeinsamen Koordinatensystem dargestellt. Hierbei wurde die TPR gegen die FPR aufgetragen. Zudem gibt die AUC über die Detektionsrate des Algorithmus beim Vorliegen verschiedener Fremdmaterialien.

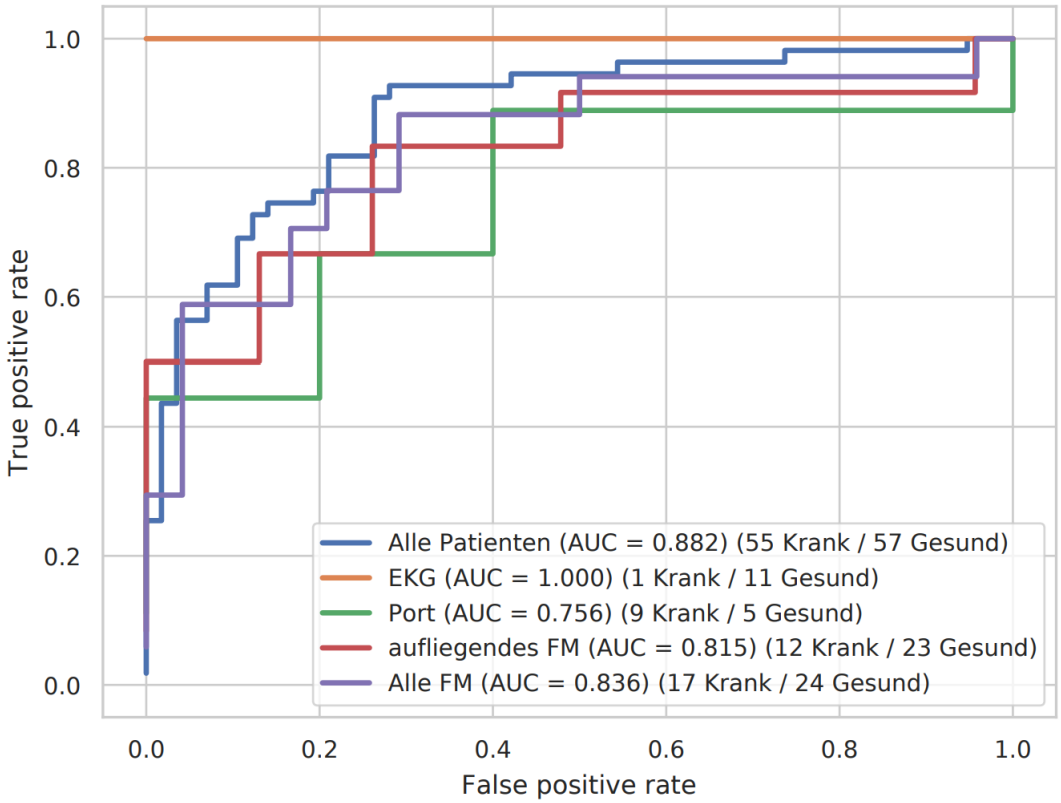


Abbildung 13: ROC-Kurve der Ergebnisse der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial

Der in Abbildung 13 genannte Begriff ‚Krank' beschreibt das Vorhandensein eines oder mehrere Lungenrundherde im Thoraxröntgenbild. ‚Gesund' beschreibt kein Vorhandensein von einem oder mehreren Lungenrundherden im Thoraxröntgenbild.

Ergebnisse der Detektion von Lungenrundherden in Thoraxröntgenbildern aller untersuchten Patienten des Datensatzes dargestellt in einer ROC-Kurve

Hierbei wurden 55 Thoraxröntgenbilder mit Vorkommen von Lungenrundherden und 57 Thoraxröntgenbilder ohne das Vorliegen von Lungenrundherden eingeschlossen. Die 112 Thoraxröntgenbilder mit ihrer Verteilung innerhalb der einzelnen Unterkategorien wurden in Kapitel 4.3.1 in Tabelle 10 aufgeführt. Die hier beschriebene ROC-Kurve stellte damit eine Gesamtübersicht über die Detektionsleistung des Algorithmus dar. Die Vielfalt der berücksichtigten Thoraxröntgenbilder und die daraus resultierende ROC-Kurve repräsentierte hier am ehesten die Detektionsleistung des Algorithmus im Alltag. Die AUC dieser Kurve betrug 0,882 und war durch die große Anzahl an berücksichtigten Daten als sehr valide einzuschätzen.

Ergebnisse der Detektion von Lungenrundherden unter Berücksichtigung des Vorkommens von EKG-Elektroden

Im folgenden Abschnitt werden die Ergebnisse der Detektion von Lungenrundherden unter Berücksichtigung des Vorkommens von EKG-Elektroden in einer ROC-Kurve aufgeführt. Hierbei wurden insgesamt 12 Thoraxröntgenbilder berücksichtigt. Ein Patient enthielt Thoraxröntgenbilder mit Lungenrundherden und Fremdmaterial im Sinne von EKG-Elektroden. 11 Patienten, die keine Lungenrundherde, aber EKG-Elektroden im Thoraxröntgenbild aufwiesen, stellten die gesunde Kontrolle dar. Die ROC-Kurve verläuft parallel zur Abszisse auf Höhe der y-Achse bei einem Wert von y = 1. Ein Wert von 1 auf der y-Achse sagt aus, dass alle vorliegenden Lungenrundherde im Röntgenbild auch als solche vom Algorithmus detektiert wurden. Zudem beschreibt die Kurve, dass der Algorithmus des maschinellen Lernens zwischen Lungenrundherden und EKG-Elektroden vollständig differenzieren konnte. Die AUC wies demnach einen Wert von 1 auf und zeigte, dass der Algorithmus die Lungenrundherde zu 100 % identifizierte und keine Fehldiagnose seitens des Algorithmus durch das vorhandene Fremdmaterial getroffen wurde. Aufgrund der geringen Anzahl an Thoraxröntgenbildern in diesem Experiment (1 Krank/11 Gesund) sollte dieses Ergebnis jedoch mit Vorsicht betrachtet werden.

Ergebnisse der Detektion von Lungenrundherden unter Berücksichtigung des Vorkommens eines implantierten Portsystems

Im folgenden Abschnitt wird die Detektion von Lungenrundherden durch den Algorithmus bei gleichzeitigem Vorliegen eines implantierten Portsystems auf dem Thoraxröntgenbild untersucht. Insgesamt wurden 14 Thoraxröntgenbilder berücksichtigt, von denen 9 Lungenrundherde enthielten und 5 die gesunde Kontrolle mit einem implantierten Portsystem darstellten.

Optisch fällt auf, dass diese Kurve, im Vergleich zu allen anderen ROC-Kurven im Koordinatensystem, der Winkelhalbierenden am nächsten liegt. Nach der ROC-Kurve der EKG-Elektroden zeigt diese Kurve die zweit geringste Anzahl an berücksichtigten Wertepaaren. Die AUC zeigte einen Wert von 0,756 auf und beschrieb somit den geringsten Wert, der hier abgebildeten ROC-Kurven.

Ergebnisse der Detektion von Lungenrundherden unter Berücksichtigung des Vorkommens von aufliegendem Fremdmaterial

Im Folgenden wird analysiert, inwiefern das Vorkommen von aufliegendem Fremdmaterial die Leistung des Algorithmus bei seiner Detektion von Lungenrundherden beeinflusst. 12 von 35 berücksichtigten Patienten enthielten Lungenrundherde. Somit stellten 23 Patienten und deren Thoraxröntgenbilder die gesunde Kontrolle dar. Die AUC dieser ROC-Kurve betrug 0,815 und zeigte nach der ROC-Kurve die, die implantierten Portsysteme darstellte, den geringsten Wert an.

Ergebnisse der Detektion von Lungenrundherden unter Berücksichtigung des allgemeinen Vorkommens von Fremdmaterial

In diesem Absatz werden die Ergebnisse der Detektion von Lungenrundherde unter dem Einfluss des allgemeinen Vorkommens von Fremdmaterialien durch den Algorithmus des maschinellen Lernens zusammengefasst dargestellt. Die Kurve gibt eine allgemeine Aussage über die Funktion des Algorithmus beim Vorliegen verschiedener Fremdmaterialien, wie EKG-Elektrode, implantierte Portsysteme und aufliegendem Fremdmaterial. Eingeschlossen wurden hierbei 24 Thoraxröntgenbilder ohne vorliegende Lungenrundherde und 17 Thoraxröntgenbilder mit vorliegenden Lungenrundherden. Die ROC-Kurve erreichte die drittgrößte AUC von 0,836 in dieser Abbildung.

6 Diskussion

Ziel des sechsten Kapitels ist es, die Ergebnisse unter Berücksichtigung der Fehlerquellen und im Rahmen der erarbeiteten Methodik zur Identifikation von Lungenrundherden einzuordnen und hinsichtlich ihrer Eignung in der Praxis zu beurteilen. In Kapitel 6.1 wird die Methodik dieser Arbeit diskutiert. Dies umfasst die Fehlerquellen, die beispielsweise durch technische Voraussetzungen oder Unterschiede in der Befundung zu Stande kommen können. Kapitel 6.2 befasst sich anschließend mit der Diskussion der Ergebnisse der Reader Studie, als auch mit der Diskussion über die Ergebnisse der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial.

6.1 Diskussion der Methoden

Im Folgenden werden möglichen Fehlerquellen aufgeführt und deren potentielle Auswirkung auf die Ergebnisse erläutert. Die Fehlerquellen lassen sich in mehrere Kategorien, wie technische Voraussetzung und Befundung, unterscheiden.

Technische Voraussetzungen

Die Zeitspanne der eingeschlossenen Thoraxröntgenaufnahmen und die damit verbundenen leicht unterschiedlichen technischen Voraussetzungen macht es nicht möglich, alle Aufnahmen unter den gleichen Rahmenbedingungen zu evaluieren.

Die Verwendung von Röntgengeräten verschiedener Baujahre und Hersteller kann zu Unterschieden hinsichtlich der Bildqualität und nachgelagerten Auswertbarkeit führen. Allgemein gilt, dass Schärfe und Kontrast des Bildes die Bildqualität der Röntgenaufnahme wesentlich bestimmen. Die Bildschärfe nimmt mit zunehmendem Abstand zwischen Röntgendetektor und darzustellendem Objekt ab. Zugleich ist die Bildqualität abhängig von der Strahlendosis, dem Filter und den Maßnahmen zur Streustrahlenreduktion. Streustrahlung entsteht durch das Auftreffen von Röntgenstrahlen auf das Gewebe, an welchem sie zum Teil abgelenkt werden. Hierbei entstehen Strahlen, die beispielsweise schräg auf den Röntgenfilm treffen und somit nicht zur anatomischen Abbildung der Strukturen dienen. Durch ein Raster, das in Strahlengangsrichtung hinter dem Objekt positioniert wird, besteht die Möglichkeit, die Streustrahlung zu reduzieren. (Albes, Lippek, Müller, & Weier, 2017) Inwieweit diese Vorgänge auf die Gesamtheit der Thoraxröntgenbilder gleichermaßen angewendet wurden, lässt sich rückwirkend nicht überprüfen. In dem Zusammenhang muss davon ausgegangen werden, dass die Thoraxröntgenbilder eine unterschiedliche Bildschärfe, Kontraste und somit Bildqualität aufweisen. Zudem sind die zum Training des Algorithmus verwendeten Thoraxröntgenbilder auf eine Pixelgröße von 512x512 angepasst, um die Rechenleistung zu reduzieren und einen einheitlichen Standard zu erstellen. Diese Modifizierungen können die Qualität der Befundung durch den Algorithmus reduzieren und

entsprechen möglicherweise somit nicht dem klinischen Arbeitsumfeld. Dem gilt es gegenüberzustellen, dass die Radiologen mit Thoraxröntgenbildern mit der vollen Auflösung arbeiteten und somit auch die Annotierung von Lungenrundherden unter diesen Verhältnissen vorgenommen wurde.

Befundung

Ein weiteres wichtiges Kriterium stellt die Befundung der Thoraxröntgenbilder dar. Diese mögliche Fehlerquelle muss im Verlauf der Arbeit in unterschiedlichen Stadien berücksichtigt werden.

Wie bereits in Kapitel 4.2 beschrieben, filtert das informationstechnische Systems (Abk. IT-System) des Krankenhauses, anhand der Stichwörter, die im Befund enthalten sind, die zutreffenden Patienten und Thoraxröntgenbilder heraus. Ausschlaggebend für diese Extrahierung dieser Patientendaten ist der Befundbericht. Dieser wird von zwei Ärzten (im Mittel drei und elf Jahre Berufserfahrung) erstellt. Durch dieses mehrstufige Verfahren mit einer hohen beruflichen Expertise in der Diagnostik kann die Beurteilung bzgl. Lungenrundherde im Thoraxröntgenbild als sehr valide gewertet werden. Die Sortierung der Patientendaten in die unterschiedlichen Kategorien ist von einer Doktorandin der Medizin im 9. Fachsemester durchgeführt worden. Diese hatte sowohl den Befundbericht als auch die Röntgenaufnahmen und weitere Diagnostik wie CT-Bilder, Biopsieergebnisse und Verlaufskontrollen vorliegen und konnte somit die Zuordnung in die unterschiedlichen Kategorien vornehmen. Durch diesen Prozess wurde die Qualität der Grundwahrheit, anhand dessen der Algorithmus trainiert und getestet wird, festgelegt. Zu berücksichtigen gilt jedoch, dass die Grundinformationen, die sich aus Befundbericht, CT-Bild und Biopsie ergeben, nicht immer vollständig gesichert sind. Denn bei der erstmaligen Kategorisierung der Röntgenbilder, basierend auf dem dazugehörigen Befund und der Darstellung der Lungenrundherde im Thoraxröntgenbild wurden viele Lungenrundherde als auffällig, suspekt oder abklärungsbedürftig betitelt. Nicht alle dieser Auffälligkeiten konnten im weiteren diagnostischen Verlauf bestätigt oder widerlegt werden. Bei der Auswahl der Bilder wurde darauf besonderen Wert gelegt, dass den Thoraxröntgenbildern mit suspekten Auffälligkeiten eine CT-Aufnahme folgte. Die CT-Aufnahme kann Aufschluss über den Verdacht der Malignität der Auffälligkeiten geben. Eingeschlossen wurden ausschließlich Thoraxröntgenbilder mit Lungenrundherden, bei denen das Vorliegen der Rundherde durch eine CT-Aufnahme gesichert worden war.

Eine Schwierigkeit bei der Annotation der Bilder ergab sich durch das hauseigene Annotationsprogramm. Die Annotierung der Lungenrundherde konnte nur an einem normalen Monitor vorgenommen werden. Somit war eine Bewertung der Lungenrundherde über die radiologischen Monitore und eine hochauflösende Fensterung der Thoraxröntgenbilder nicht möglich. Insbesondere unterdurchschnittlich kleine Lungen-

rundherde konnten unter diesen technischen Gegebenheiten nicht eindeutig als suspekt und auffällig deklariert werden.
Sowohl die Thoraxröntgenbilder, bei denen die Anonymisierung zu Problemen bei der Annotation von Lungenrundherden führte, als auch die, bei denen die technischen Gegebenheiten Schwierigkeiten hervorriefen, wurden aus dem Datensatz entfernt und waren somit nicht Teil der Studie.
Die händische Annotierung der Lungenrundherde wurde durch einen Radiologen im dritten Berufsjahr ausgeführt. Hierbei bestand die Option des Zooms in das Thoraxröntgenbild, sodass die Möglichkeit gegeben war alle Lungenrundherde möglichst rand- und pixelgenau zu markieren. Dem Annotationsprogramm war es jedoch nicht möglich, einen standardisierten Zoom für alle zu markierenden Thoraxröntgenbilder festzulegen. Das bedeutet, dass das unterschiedliche und subjektive Zoomverhalten durch den Radiologen, bezogen auf die verschiedenen Thoraxröntgenbilder, Abweichungen in der Genauigkeit der Pixel-Markierungen aufweisen kann. Dadurch können die Segmentierungen in unterschiedliche Qualitätsstufen erstellt worden sein. Die Qualität hängt sehr von der gewählten Pinselgröße zur Markierung der Lungenrundherde und Zoomstufe ab. Eine große Pinselgröße und niedrige Zoomstufe impliziert etwa eine ungenauere Segmentierung.
In der Reader Studie selbst war es den Radiologen nicht möglich die Fensterfunktion zu verwenden. Somit wurde die Annotierung der Lungenrundherde nur aus einer Perspektive ausgeführt, was zu einer Standardisierung der Studie führte. Kritisch zu hinterfragen ist der tatsächliche Nutzen einer Fensterfunktion mit einer daraus möglicherweise verbesserten Detektionsleistung bei nur 10 Sekunden Befundungszeit pro Thoraxröntgenbild innerhalb der Reader Studie. Dem anzufügen ist, dass in der Reader Studie nur posterior-anteriore Thoraxröntgenbilder für den Algorithmus und die Radiologen zur Verfügung standen. Durch Hinzunahme seitlicher Ansichten könnte man die Genauigkeit der Detektion verbessern, jedoch würde sich hiermit auch die Befundungszeit durch den Algorithmus verdoppeln. Anzufügen gilt es, dass die seitlichen Thoraröntgenbilder ebenfalls vorab durch einen Experten segmentiert werden müssen.
Darüber hinaus stammt der Großteil der Daten aus dem PACS des Klinikum rechts der Isar und der JSRT Datenbank. Aus diesem Grund kommen Thoraxröntgenbilder bestimmter Geräte oder Bevölkerungsgruppen nicht in dem in dieser Arbeit verwendeten Datensatz vor und sind somit unterrepräsentiert. Beispielsweise sind Thoraxröntgenbilder von Kindern und Säuglingen in dieser Arbeit nicht mit eingeschlossen worden. Dieser Aspekt könnte es erschweren das Modell auf verschiedene Bevölkerungsgruppen und Geräte zu übertragen. (Zech et al., 2018) Dennoch könnte das Problem mittels einer Erweiterung des Datensatzes hinsichtlich dieser Problematik gelöst werden. Zudem gibt es die Möglichkeit über öffentliche Plattformen wie bspw. engl. The Cancer Imaging Archive (Abk. TCIA) auf große, internationale und geprüfte Daten-

sätze zuzugreifen. Somit würden nicht nur mögliche Unterrepräsentationen verhindert werden, sondern zugleich könnten auch größere Mengen an Daten auf das Training und die Testung des Algorithmus angewendet werden können. ("The Cancer Imaging Archive," 2021)

6.2 Diskussion der Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse der Reader Studie, als auch die Ergebnisse der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial durch den Algorithmus diskutiert. Die Fragestellung, mit der sich die Reader Studie beschäftigt, ist die Untersuchung, ob Modelle des maschinellen Lernens potentiell zu einer sicheren Diagnose durch CADe-Systeme beitragen können.

Diskussion der Ergebnisse der Reader Studie

Die in Kapitel 5.4, Tabelle 18, dargestellten Ergebnisse der Reader Studie zeigen, dass der Algorithmus im Vergleich zu den beiden Radiologen eine ähnliche Detektionsleistung aufweist. Der Algorithmus erzielt die gleiche Anzahl an TP und FN, wie Radiologe B. Lediglich in der Anzahl der FP unterscheiden sich diese voneinander. In diesem Zusammenhang ist zu erwähnen, dass einige FP durch sich überlappende oder gegenseitig einschließende Begrenzungsrahmen zustande kommen. Würden diese Boxen nicht einzeln gezählt werden, so könnte die FPR des Algorithmus in einigen Fällen deutlich niedriger sein, als hier dargestellt. Die Differenz der FPR erklärt die unterschiedlichen F2 Scores. Darüber hinaus lässt die hohe Empfindlichkeit des Algorithmus trotz der großen Anzahl von FPs darauf schließen, dass im Prinzip mehr Lungenrundherde nachweisbar wären, was bedeutet, dass das Modell durch Training mit mehr Daten verbessert werden könnte. Zudem sollte berücksichtigt werden, dass die Bewertung des Algorithmus mittels der verschiedenen Parameter, die zur Evaluation dienen, abhängig ist. Das bedeutet, dass es möglicherweise andere bzw. bessere Auswertungsparameter für die bestmögliche Detektionsleistung gibt, die in dieser Modellanalyse nicht berücksichtigt werden können. (Schultheiss et al., 2020) (Schober, 2019)

Bestehende CAD-Systeme dienen derzeit der Unterstützung der Radiologen im klinischen Alltag. Die hohe Empfindlichkeit bei der Detektion von Lungenrundherden bedeutet zugleich, dass diese viele FP erkennen und somit Radiologen auf potentielle Lungenrundherde aufmerksam machen können. (Bush, 2016) Auch in dieser Arbeit zeigt sich eine höhere Rate an FP durch den Algorithmus, verglichen mit den Radiologen. Ein falsch hoher FP Wert durch den Algorithmus wird unter Annahme des Verfahrens des ‚second-look' durch den Radiologen relativiert. Im sogenannten ‚second-look' sortiert der Radiologe anschließend die FP Detektionen nach ihrer Relevanz aus. Studien haben jedoch gezeigt, dass es schwierig für Radiologen ist, wahre

Lungenrundherde effektiv von FP Detektionen durch den Algorithmus zu unterscheiden. Somit wäre es wünschenswert, dass CAD-Systeme die Anzahl an FP Detektionen durch verbesserte Funktion von vorneherein reduzieren könnten. Zudem sind die CAD-Systeme derzeit noch nicht in der Lage, mit einer 100%igen Sicherheit alle Lungenrundherde auf einem Thoraxröntgenbild zu identifizieren, sodass es immer noch die Aufgabe des Radiologen ist, das gesamte Bild auf das Vorkommen von Lungenrundherden zu untersuchen. (Bush, 2016) Anzufügen gilt es, dass in dieser Arbeit kein Fokus auf die Aktivität und Dignität der Lungenrundherde gesetzt wurde und die Einschätzung dessen derzeit ebenso Aufgabe des Radiologen ist. Dennoch zeigt die Reader Studie, dass der Algorithmus 7 von 15 Lungenrundherde detektierte, die kein Radiologe in der Reader Studie identifizierte, und unterstreicht somit die Aussage, dass CAD-Systeme diagnostisch positiv unterstützend wirken können. Zudem ermöglichen Algorithmen des maschinellen Lernens einen objektiven Vergleich von im Verlauf entstandenen Thoraxröntgenbildern in Hinsicht auf Vorkommen, sowie Größen- und Formunterschiede von Lungenrundherden (Dreher et al., 2019).

Aus Kapitel 5.3 ergibt sich, dass Mamillenschatten, trotz ihrer Form und Lokalisation im Thoraxröntgenbild, die Detektion von Lungenrundherden durch den Algorithmus nicht beeinflussen. Allgemein zeigt sich in der Auswertung jedoch, dass das Vorhandensein von Fremdmaterial die FPR von Lungenrundherden durch den Algorithmus verändert. Zunächst wäre denkbar, dass diese Beobachtung auf die automatischen Einstellungsmodi der Röntgengeräte bei gleichzeitigem Vorkommen von Fremdmaterial zurückzuführen ist. Dennoch zeigen sich deutliche Unterschiede zwischen den Unterkategorien Rundherd + Fremdmaterial und Unauffällig + Fremdmaterial. Diese Unterschiede unterstützen somit nicht die zuvor getätigte Vermutung. Mit Ausnahme der Unterkategorie Unauffällig + Fremdmaterial zeigt der Algorithmus in den übrigen Kategorien ähnliche Ergebnisse wie die beiden Radiologen.

Wie bereits in Kapitel 3 kurz erläutert, entwickelten RAJPURKAR ET AL. einen Deep-Learning-Algorithmus der 14 klinisch wichtige Pathologien in Thoraxröntgenbildern erkennt und die Erkrankung lokalisieren kann. Der Algorithmus wurde an 9 praktizierenden Radiologen anhand eines Validierungssatzes von 420 Bildern evaluiert. Die Grundwahrheit wurde durch 3 gleiche Ergebnisse von Radiologen, die auf Thoraxröntgenbilder spezialisiert sind, gebildet. Der Algorithmus erreichte bei allen 10 Pathologien eine Leistung, die vergleichbar mit den praktizierenden Radiologen ist. Der Algorithmus zeigte bei einer Pathologie eine bessere, und bei 3 Pathologien eine schlechtere Leistung als diese. Diese Studie arbeitete ebenso die zeitliche Differenz zwischen Radiologen und Algorithmus heraus und zeigte, dass die Radiologen durchschnittlich 240 Minuten und der Algorithmus 1,5 Minuten für die Evaluation von 420 Bilder benötigten. Diese Studie zeigt demnach ähnliche Ergebnisse, die auch in dieser Arbeit festgestellt werden konnten. Dennoch zeigt die in dieser Arbeit durchgeführte Reader Studie Besonderheiten, wie die Abbildung von Mamillenschatten auf Thoraxröntgenbildern, die

keine Schwierigkeit für den Algorithmus darstellten und differenziert bewertet werden konnten. Ebenso wurde in dieser Reader Studie das Verhalten des Algorithmus bei Vorkommen von Fremdmaterial auf Thoraxröntgenbildern untersucht, welches bei RAJPURKAR ET. AL. nicht berücksichtigt wurde. (Rajpurkar et al., 2018)

Dass CAD-Systeme im klinischen Alltag der Radiologen als sinnvolle Unterstützung gelten, wird in vielen Forschungsansätzen dargestellt. Ergebnisse zeigen, dass der Radiologe meist nicht durch ein CAD-System zu ersetzen ist. Durch den Einsatz von Algorithmen des maschinellen Lernes können jedoch meist bessere Werte bei der Detektion von Lungenrundherden erzielt werden, als wenn zwei Radiologen nach dem Prinzip des „double-proof-reading“ verfahren. (Rubin et al., 2005) Zudem ist die Leistung der CAD-Systeme objektivierbar, nahezu unabhängig vom derzeitigen Ärztemangel in den Krankenhäusern (Mintz, 2019) und effizienter in Betrachtung von Zeit- und Kostenersparnis (Alkadhi et al., 2012) im Vergleich zu Radiologen. Dem gilt anzufügen, dass die Leistung des Untersuchers abhängig von dessen Erfahrung und Ermüdbarkeit ist (Mintz, 2019). Mittels CAD-Systeme kann eine flächendeckende Verfügbarkeit, auch in einkommensschwachen Ländern, gewährleistet werden, sofern die technischen Voraussetzungen erfüllt sind (Ngoya et al., 2016b).

Nach Berichten der Weltgesundheitsorganisation (engl. World Health Organization) haben mehr als 4 Milliarden Menschen keinen Zugang zu medizinischer Bildgebung. Selbst in Industrieländern mit fortschrittlichen Gesundheitssystemen könnte ein automatisiertes System zur Interpretation von Thoraxröntgenbildern einen immensen Nutzen bringen. (Welling et al., 2011) Dies beschreibt die Notwendigkeit einer weltweiten Vergrößerung der Ressourcen, die durch ein CAD-System, wie es in dieser Arbeit beschrieben wird, verbessert werden könnte.

Es konnte hier gezeigt werden, dass unter 15 FN Detektionen beider Radiologen je 7 und 6 Lungenrundherde durch den Algorithmus gekennzeichnet wurden. Von den 6 Lungenrundherden, die von keinem Radiologen gefunden wurden, wären vom Algorithmus mindestens 2 Lungenrundherde nachgewiesen worden. Dies führt zu der Schlussfolgerung, dass die radiologische Diagnostik im Rahmen von Screening von Thoraxröntgenbildern stark von solchen Algorithmen profitieren könnte. Eine zusätzliche Anforderung für den Einsatz dieser Techniken ist jedoch eine Verbesserung der Spezifität der Modelle. Bisher zeigt der Algorithmus eine relativ hohe Anzahl FN Detektionen, die in einer klinischen Umgebung dazu führen würde, dass die Radiologen wiederum Zeit aufwenden müssten, um diese in der Realität zu bewerten. (Schultheiss et al., 2020) (Schober, 2019) Diese Situation gilt es dann differenziert, zwischen der Zeit der manuellen Befundung durch den Radiologen selbst, im Vergleich zur Zeit der Überprüfung FN Markierungen im Thoraxröntgenbild, und dessen präzise Evaluation FN Detektionen, zu betrachten. Hiermit zeigt sich das Verbesserungspotential der algorithmischen Leistung in Hinsicht auf die Reduktion der FP Detektionen. Somit wäre

es ebenfalls denkbar, CAD-Systeme in Screening-Untersuchungen einzuführen und die Ergebnisse anschließend durch einen ausgebildeten Arzt verifizieren zu lassen.

Zusammenfassend konnte gezeigt werden, dass der Algorithmus des maschinellen Lernens eine ähnlich gute Leistung wie die Radiologen erzielen kann. Um die Zuverlässigkeit in Hinsicht auf die klinische und effiziente Anwendung von CAD-Systemen zu erhöhen, sind Verbesserungen in der Empfindlichkeit und Reduktion FP Detektionen erforderlich. In dieser Arbeit, als auch anhand anderer Studien wurde die Relevanz und Anwendbarkeit von CAD-Systemen aufgezeigt.
In dieser Studie gilt es als kritisch zu betrachten, dass ausschließlich maligne Lungenrundherde eingeschlossen wurden. Das bedeutet, dass der Algortihmus nur auf die Erkennung maligner Lungenrundherde trainiert wurde. Somit wurde die Detektion von gutartigen Lungenrundherde nicht eingeschlossen, und die Unterscheidung zwischen gut- und bösartigen Lungenrundherden wurde nicht trainiert. Zudem wurde der Algorithmus möglicherweise für kleine Lungenrundherde (dh. kleiner als 1 cm Durchmesser) untertrainiert. Nicht genauer untersucht wurden die Lungenrundherde, die nicht vom Algorithmus detektiert wurden und somit als FN Detektion klassifiziert wurden. Es ist denkbar, dass es sich bei diesen FN Detektionen um Lungenrundherde handelt, die im Vergleich deutliche Größenabweichungen zu TP Detektionen aufweisen. Desweiteren wurden nicht alle im klinischen Alltag vorkommenden häufigen Lungenerkrankungen in den Thoraxröntgenbildern der Kategorien Unauffällig + und Rundherd + abgebildet. Darüber hinaus haben wir uns in dieser Studie nur mit Röntgenaufnahmen des Brustkorbs befasst, die mittels posterior-anterioren Projektionen aufgenommen wurden. Möglichkeiten, welche die Leistung der angewandten Deep-Learning-Technologie weiter verbessern könnten, sind zum Beispiel die Verwendung von Bildern der Seitansicht, Bildern mit Knochenunterdrückung/Knochenansicht, sowie die Verwendung von Doppel-Energie-Röntgen-Absorptiometrie. Zu berücksichtigen gilt außerdem, dass diese Arbeit eine retrospektive Studie, die nicht exakt die reale Umgebung darstellten konnte, und ausschließlich vom medizinischen Personal des Klinikum rechts der Isar befundet wurde. Weitere Untersuchungen sind erforderlich, um die Anwendbarkeit diesen Algorithmus in einer prospektiven multi-institutionellen Studie zu bestimmen, oder diesen bei der Analyse prospektiver multi-institutioneller Datensätze zu bewerten.

Diskussion der Ergebnisse der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial

In Kapitel 5.5 sind die Ergebnisse der Detektion von Lungenrundherden unter dem Einfluss von Fremdmaterial graphisch und schriftlich erläutert. Diese werden im zweiten Teil dieses Abschnitts diskutiert und hierbei die Möglichkeit der Anwendung von Algorithmen des maschinellen Lernens im klinischen Alltag erläutert. Vorab werden

jedoch die Ergebnisse der Fremdmaterial-enthaltenden Unterkategorien der Reader Studie aus Kapitel 5.4 kritisch hinterfragt und Lösungsansätze zur Verbesserung der Ergebnisse diskutiert.

Im Folgenden werden die Ergebnisse der Unterkategorien U + FM und RH + FM aus Tabelle 19, die die Kategorien-spezifischen Ergebnisse aus der Reader Studie darstellt, diskutiert.

Der Algorithmus identifiziert in der Kategorie der Rundherde 16 TP und 13 FN Detektionen. Diese Detektionsraten entsprechen derer des best-evaluierten Radiologen der Reader Studie (Radiologe A).

Bei Betrachtung der Werte der Unterkategorien U und RH fällt auf, dass der Algorithmus in der Kategorie der unauffälligen Thoraxröntgenbilder mehr FP Detektionen aufweist, verglichen zur Kategorie der Rundherde. Auffällig ist diese Abweichung im Vergleich zu beiden Radiologen.

Zu berücksichtigen gilt jedoch, dass die Unterkategorie U 3 Thoraxröntgenbilder mehr als die Kategorie RH im Testdatensatz enthält. Ebenso kritisch zu hinterfragen ist, ob es sinnvoll gewesen wäre, den Algorithmus mit einer gleichen Anzahl von unauffälligen und Rundherd-enthaltenen Thoraxröntgenbildern zu trainieren, um somit seine FPR im Bereich der unauffälligen Bilder ggf. reduzieren zu können. Denkbar ist, dass größere Datensätze mit ausgewogenen Anteilen von Thoraxröntgenbildern mit und ohne Lungenrundherde die Leistung des Algorithmus verbessern können. Mit dem Fokus überwiegend Thoraxröntgenbilder mit Lungenrundherden als Trainings-Datensatz zu verwenden wird der Algorithmus auf die Detektion von Lungenrundherden sensitiviert. Um Algorithmen im klinischen Alltag effizient nutzen und diese als zuverlässig beschreiben zu können, ist es notwendig, dass die FPR reduziert wird. Eine hohe FPR durch den Algorithmus erfordert, dass Radiologen die FP Detektionen erneut bewerten müssen. Dennoch zeigen einige Studien, dass genau dieser Vorgang zu verbesserten Ergebnissen in der Diagnosestellung führt und eine Unterstützung durch den Algorithmus und seine erhöhte FPR positive Effekte auf die Detektion von Lungenrundherde zeigt (Monnier-Cholley et al., 1998). Weiter zu erforschen gilt es, warum der Algorithmus in der Kategorie RH + FM weniger FP Detektionen aufweist, verglichen zur Kategorie U + FM, oder ob sich diese Beobachtung mit zunehmender Anzahl an Daten verändert. Ein wichtiges Ergebnis ist allerdings, dass ein gut trainiertes Model des maschinellen Lernens nur 2 FP Detektionen von Lungenrundherden in insgesamt 59 Thoraxröntgenbildern mit Fremdmaterial detektierte, wie wir es in SCHULTHEISS ET AL. publiziert haben. Neben dem Potenzial von maschinellem Lernen zeigt unsere Studie auch, dass vor der klinischen Verwendung eines solchen Systems sorgfältig geprüft werden muss, welche Bildmerkmale zu falschen Entscheidungen von Algorithmen des maschinellen Lernens beitragen können und somit Limitierungen für die Anwendung dieser sein können. (Schultheiss et al., 2020)

Die ROC-Kurve aus Abbildung 13 berücksichtigt alle auffälligen als auch unauffälligen Thoraxröntgenbilder und zeigt eine AUC von 0,882. Dieser Wert ist wie bereits in Kapitel 5.5 erläutert, der zweithöchste Wert, verglichen zu den anderen untersuchten Kategorien. Anzumerken gilt hier, dass dieser Datensatz aus 55 Thoraxröntgenbilder mit Lungenrundherden und 57 Thoraxröntgenbildern ohne Lungenrundherde besteht und somit die Thoraxröntgenbilder innerhalb ihrer dieser Kategorien nicht gleichverteilt sind. Dennoch stellt dieser Datensatz von seiner Zusammensetzung und Vielfalt der Unterkategorien, eingeschlossen den verschiedenen Fremdmaterialien, einen für den klinischen Alltag repräsentativen Datensatz dar und zeigt einen guten Wert der AUC. Die Analyse und die Beobachtung des Algorithmus bei gleichzeitigem Vorkommen von Fremdmaterial sind im klinischen Alltag von großer klinischer Relevanz und derzeit noch nicht allumfänglich erforscht.
Viele Patienten zeigen auf ihren Thoraxröntgenbildern das Vorhandensein von rundlichen EKG-Elektroden auf der Brustwand. Dennoch ist die Leistung des Algorithmus in dieser Arbeit nur anhand von 12 Thoraxröntgenbildern überprüft worden. In Abhängigkeit von der Form des Fremdmaterials hätte man davon ausgehen können, dass EKG-Elektroden die Leistung des Algorithmus verändern. Die ROC-Kurve, die die Leistung des Algorithmus bei gleichzeitigem Vorkommen von EKG-Elektroden darstellt, zeigt dass dieser alle Thoraxröntgenbilder mit EKG-Elektroden richtig analysiert hat. Kritisch zu betrachten ist hier allerdings die geringe Anzahl an Daten und die Berücksichtigung von nur einem Thoraxröntgenbild mit gleichzeitigem Vorliegen von Lungenrundherden und EKG-Elektroden. Diese Leistung des Algorithmus lässt sich mit zunehmender Anzahl an Daten überprüfen und ggf. verifizieren und ist von großer Bedeutung, da ein EKG eine häufige Untersuchung in der ambulanten als auch stationären Diagnostik ist.
Ebenso als klinisch relevant zu bewerten sind implantierte Portsysteme, die sehr häufig bei onkologische Patienten mit Chemotherapie Anwendung finden. In der Regel wird zunächst nach Implantation des Portsystems dessen Lage mittels Thoraxröntgenbild kontrolliert. Von größerer Bedeutung ist jedoch die regelmäßige Thoraxröntgenbildgebung mit der Frage nach Lungenrundherden innerhalb dieses Patientenkollektivs. In dieser Arbeit wird die Funktion des Algorithmus bei Vorliegen von Portsystemen anhand von 9 Thoraxröntgenbildern mit Lungenrundherden und 5 Thoraxröntgenbilder ohne Lungenrundherde überprüft. Die Untersuchung dieser Kategorie zeigt als einzige Gruppe einen größeren Anteil an Thoraxröntgenbildern mit Lungenrundherden als ohne Lungenrundherde. Der Algorithmus zeigt hier im Vergleich zu den anderen Kategorien die AUC mit dem geringsten Wert. Diese könnte einerseits durch die Größe des Ports zu erklären sein, da dieser möglicherweise Lungenrundherde teilweise oder gar ganz in der posterior-anterioren Aufnahme verdeckt. Zum anderen ist denkbar, dass eine Überlagerung durch den Portschlauch, der das Thoraxröntgenbild durchzieht, die Detektionsleistung abschwächt. Auch hier würde ein größerer

Trainingsdatensatz und besonders die Kombination mit seitlichen Ansichten des Thorax die Leistung des Algorithmus verbessern. So könnten vom Portsystem verdeckte Bereiche reduziert und die Detektionsrate durch den Algorithmus gesteigert werden. Gleiches Problem der Verdeckung und Überlagerung von Fremdmaterial auf Thoraxröntgenbildern zeigt sich bei der Kategorie, die aufliegendes Fremdmaterial aufweist. Manchen Patienten ist es nicht möglich ihnen aufliegendes Fremdmaterial, wie beispielsweise Piercings oder Schmuck, für eine Thoraxröntgenbildaufnahme zu entfernen. Zudem belassen einige Patienten Ohrringe oder Halsketten während der Röntgenaufnahme an ihrem Körper, da diese sich nicht direkt auf die Lungenfelder projizieren und für die Bewertung dieser somit irrelevant sind. Inwiefern aufliegende Fremdmaterialien die Bildqualität durch Röntgengeräte verändern und somit indirekt die Leistung des Algorithmus beeinträchtigen, ist derzeit noch nicht eindeutig geklärt. Wie sehr das unterschiedliche, aufliegende Fremdmaterial die Detektionsleistung des Algorithmus beeinflusst, kann aus diesen Daten nicht reell ermittelt werden, da nahezu $^2/_3$ der Thoraxröntgenbilder unauffällig sind. Bei den Thoraxröntgenbildern dieser Gruppe mit Lungenrundherd kann anhand der ROC-Kurve nicht abgeleitet werden, bei welcher Art von aufliegendem Fremdmaterial und welcher Lokalisation der Algorithmus Probleme hat. Eine genauere Analyse dieser Bilder wäre in Anbetracht dieser konkreteren Untersuchung durchaus von Bedeutung.

Berücksichtigt man alle Thoraxröntgenbilder mit Fremdmaterial, so zeigt sich die dritthöchste AUC aller Unterkategorien bei einer Datensatzmenge von 41 Thoraxröntgenbildern. Der Algorithmus liefert in dieser Kategorie eine ähnliche Leistung, wie in der Kategorie, die alle Thoraxröntgenbilder des gesamten Datensatzes enthält. Aufgrund der größeren Anzahl berücksichtigter Thoraxröntgenbildern mit Lungenrundherden könnte diese Beobachtung darauf schließen lassen, dass der Wert der AUC nicht abhängig vom Vorkommen unauffälliger Thoraxröntgenbilder ist. Um diese Beobachtung zu bestätigen, wäre ein ausgewogenes Verhältnis von unauffälligen und nicht unauffälligen Thoraxröntgenbildern innerhalb der Kategorien sinnvoll.

Zusammenfassend ist anzumerken, dass der für die Detektion von Fremdmaterial verwendete Testdatensatz wenig Bilder hinsichtlich der einzelnen Kategorien enthält (z. B. 12 Thoraxröntgenbilder für die Untergruppe, die EKG-Elektroden enthält). Eine größere Anzahl an Bildern könnte die Leistung des Algorithmus genauer darstellen. Die zuvor vorgestellten Ergebnisse für die Detektion von Lungenrundherden basieren auf einem Testdatensatz von 75 Thoraxröntgenbildern. Studien, dessen Algorithmen die menschliche Leistung erfüllen, oder gar übertreffen, wie beispielsweise bei NAM ET AL. verwendeten mehr als doppelt so viele Thoraxröntgenbilder innerhalb ihres Testdatensatzes (181 Thoraxröntgenbilder mit Lungenrundherden und 62 ohne Lungenrundherde) (Nam et al., 2019). In vielen Studien ist die Anzahl unauffälliger und nicht unauffälliger Thoraxröntgenbilder zwischen den einzelnen Kategorien sehr unterschiedlich, sodass ein Vergleich oder Rückschlüsse auf die Detektionsleistung hinsichtlich

dieses Vorkommens schwierig ist. Dem anzufügen gilt, dass zum Training des in dieser Arbeit verwendeten Algorithmus ausschließlich posterior-anteriore Thoraxröntgenbilder mit einer Auflösung von 512x512 Pixeln verwendet wurden. Eine verbesserte Leistung könnte durch die Hinzunahme von seitlichen Aufnahmen geschaffen werden. Abzuwägen gilt ebenso, ob die derzeitige FPR durch den Algorithmus des maschinellen Lernens zu einer Zeitersparnis für Radiologen führt, oder ob es keine, oder gar eine negative Auswirkung auf die Effizienz hat (Bush, 2016). Eindeutig ist jedoch, dass der unterstützende Einsatz von Algorithmen des maschinellen Lernens in Kombination mit Radiologen eine verbesserte Detektionsleistung erzielt. Um die Leistung des Algorithmus unter dem Einfluss von Fremdmaterial noch detaillierter zu überprüfen, wäre eine genauere Ergebnisgewinnung und Auswertung, wie es in der Reader Studie der Fall ist, von Vorteil.

7 Zusammenfassung und Ausblick

In den letzten 20 Jahren zeigte die Forschung im Bereich des maschinellen Lernens große Fortschritte. Das Potential wurde erkannt und maschinelles Lernen findet in unterschiedlichsten Branchen Anwendung, so auch in der Medizin.

Dies erklärt die große Anzahl an wissenschaftlichen Studien, die den Einsatz und oftmals auch die Vorteile von CAD-Systemen in der Thoraxradiologie beschreiben. Als ein Vorteil von CAD-Systemen ist die Effizienzsteigerung bei der Analyse von Thoraxröntgenbildern zu nennen, wodurch die Technologie in der Vergangenheit allgemein an Interesse gewinnen konnte. (Diederich, Wormanns, & Heindel, 2001). COCCIA zeigt einen exponentiellen Anstieg in der Anzahl der wissenschaftlichen Veröffentlichungen seit 1990 im Bereich des maschinellen Lernens bezogen auf die Detektion von Lungenrundherden (Coccia, 2020).

Die in der vorliegenden Studie vorgestellten Ergebnisse zeigen, dass Algorithmen des maschinellen Lernens zur Detektion von Lungenrundherden in Thoraxröntgenbildern das Potential bieten, unter realen Bedingungen eingesetzt zu werden. Zudem können diese Algorithmen Lungenrundherde, bei gleichzeitigem Vorkommen von anderen Pathologien und Fremdmaterial, im Thoraxröntgenbild differenzieren und detektieren. Die Leistung der Algorithmen ist auf dieser Ebene mit praktizierenden Radiologen vergleichbar. Zum Teil ist diese sogar sensitiver, was sich positiv auf die Detektionsrate auswirkt. Es wurden 7 von insgesamt 15 Lungenrundherden, welche durch beide, an dieser Studie teilnehmenden Radiologen übersehen wurden, vom Algorithmus als Lungenrundherde richtigerweise detektiert. Dieses besonders sensitive Verhalten des Algorithmus ist aus klinischer Perspektive von besonderem Wert, wird jedoch durch die aktuell typischerweise genutzten Auswertungsparameter nicht klar erkennbar. Es bedarf daher für die Zukunft einer geeigneteren Darstellung solcher Ergebnisse des maschinellen Lernens.

Die klinische Integration von Algorithmen des maschinellen Lernens kann bei Anwendung eine Transformation der Patientenversorgung ermöglichen, indem die Zeit bis zur Diagnose verkürzt und der Zugang zur Interpretation von Thoraxröntgenaufnahmen verbessert wird. Der Algorithmus könnte darüber hinaus zur Priorisierung von Arbeitslisten verwendet werden, sodass die erkrankten Patienten auch in Krankenhausumgebungen, in denen Radiologen nicht sofort verfügbar sind, schnellere Diagnosen und Behandlungen erhalten. Zudem sind erfahrene Radiologen immer noch menschlichen Einschränkungen unterworfen, einschließlich Erschöpfung, Wahrnehmungsverzerrung und -einschränkungen, Wissensstand und kognitiven Verzerrungen, die alle zu Fehlern führen können und durch den Einsatz von maschinellem Lernen reduziert werden könnten. (Rajpurkar et al., 2018)

Somit zeigt diese Technologie einen Fortschritt in vielen Bereichen, die Leistung, Präzision, Zeiteffizient und Kosten verbessern kann. In der Medizin könnte eine verbesserte Patientenversorgung durch frühe Detektion und Diagnose von Krankheiten, einem verbesserten Arbeitsablauf mit Reduktion medizinischer Fehler, Kosten und somit einer Verringerung von Morbidität und Mortalität herbeigeführt werden. Entgegen vieler Aussagen soll maschinelles Lernen nicht die Arbeitsplätze der Radiologen gefährden, sondern diese und die medizinische Versorgung unterstützen oder erweitern. Hierbei ist zu berücksichtigen, dass die Diagnostik und somit auch die zu befundenen Aufnahmen mit wachsender Weltbevölkerung zunehmen werden, die Anzahl der praktizierenden Ärzte dem jedoch nicht im vollen Umfang gerecht werden kann. Maschinelles Lernen kann also dazu dienen, den Workflow zu verbessern und mit wissensbasierten Entscheidungen sichere, konsistentere und quantitative Ergebnisse liefern. Somit kann das Konzept als Unterstützung dienen, wobei die endgültige Aussage dem Menschen obliegt. Zur flächendeckenden Verwendung von CAD-Systemen, basierend auf maschinellem Lernen, sind vorab jedoch noch einige Fragen zu beantworten. Eine davon ist beispielsweise der ethische Umgang mit Datentransfer. (Mintz, 2019) Zudem besteht noch Potential in der Verbesserung von CAD-Systemen, sodass diese auch gerne von den Radiologen verwendet und als sinnvoll und zuverlässig betrachtet werden. MONNIER-CHOLLEY ET AL. zeigt, dass der gemeinsame Einsatz von Radiologen und CAD-Systemen bezogen auf die Detektion von interstitiellen Lungenerkrankungen im konventionellen Röntgenbild bessere ROC-Ergebnisse liefert, als deren alleinige Leistung. (Monnier-Cholley et al., 1998) Eine andere Studie zeigt, dass die Klassifizierung und die Dignitätsbeurteilung von Lungenrundherden in konventionellen Röntgenbildern durch einen Algorithmus sogar bessere ROC-Ergebnisse liefert, als die radiologische Vergleichsgruppe (Nakamura et al., 2000). Die Strategie zur Verbesserung dieser Systeme kann durch die Erstellung öffentlich verfügbarer Datenbanken, Trainings und Validierungen, als auch mittels gemeinsamen Forschens auf übergreifenden Plattformen am ehesten ausgeschöpft werden. Radiologen spielen bei diesen Aufgaben eine Schlüsselrolle, denn sie können ihr Wissen einbringen und bei der Erstellung hochwertig kommentierter Datenbanken für Schulung und Validierung von CAD-Systemen helfen. (van Ginneken, Schaefer-Prokop, & Prokop, 2011)

Zusammenfassend zeigt sich in der vorliegenden Arbeit, dass Algorithmen des maschinellen Lernens im Bereich der Detektion von Lungenrundherden, auch unter der Anwesenheit von Fremdmaterial, zuverlässige und sehr sensitive Ergebnisse liefern. Dennoch erfordern die Ergebnisse derzeit eine Überprüfung durch den Radiologen und können somit nicht zur alleinigen Diagnosestellung verwendet werden. Die Unterstützung durch CAD-Systeme für Radiologen ist somit eine Möglichkeit effiziente und rationale Entscheidungen im Bereich der Thoraxröntgenbildgebung zu treffen. Eine Zusammenführung verschiedener realer Datenbanken und Forschungsergebnisse

könnte das Training und abschließend die Leistung von Algorithmen in Zukunft effizienter gestalten.

V Literaturverzeichnis

Achenbach, T., Vomweg, T., Heussel, C. P., Thelen, M., & Kauczo, H. U. (2003). Computerunterstützte Diagnostik in der Thoraxradiologie - aktuelle Schwerpunkte und Techniken. *Fortschr Röntgenstr, 175*, 11.

Adler, A. (2011). *Diagnostik von Lungenrundherden unter 3 cm mittels Computertomografie bei Patienten mit maligner Vorerkrankung.* Retrieved from https://epub.uni-regensburg.de/23509/2/Lungenrundherde27022012.pdf (06.08.19)

Akan, E. (2018). Was ist ein PACS? Retrieved from https://www.itz-medi.com/was-ist-ein-pacs/ (06.12.20)

Albes, G., Lippek, V., Müller, B., & Weier, C. (2017). *Facharztprüfung Radiologie* (Vol. 4). Stuttgart, New York, Delhi, Rio: Thieme Verlagsgruppe.

Alkadhi, H., Frauenfelder, T., & Schmidt, D. (2012). Radiologie: Lungenscreening mit der Computertomographie. *Swiss Medical Forum, 12*, 25-26.

Becker, R. (2019). Convolutional Neural Networks – Aufbau, Funktion und Anwendungsgebiete Retrieved from https://jaai.de/convolutional-neural-networks-cnn-aufbau-funktion-und-anwendungsgebiete-1691/ (03.02.21)

Berger, W. G., Erly, W. K., Krupinski, E. A., Standen, J. R., & Stern, R. G. . (2001). The Solitary Pulmonary Nodule on Chest Radiography: Can We Really Tell If the Nodule Is Calcified? *American Journal of Roentgenology,*. doi:10.2214/ajr.176.1.1760201

Bergmann, T., Bölükbas, S., Beqiri, S., Trainer, S., & Schirren, J. (2007). Der solitäre Lungenrundherd, Bewertung und Therapie. *Chirurg, 8*(8), 11. doi:10.1007/s00104-007-1379-4

Bishop, C. M. (2006). *Pattern Recognition And Machine Learning*.

Bley, T. A., Baumann, T., Saueressig, U., Pache, G., Treier, M., Schaefer, O., . . . Kotter, E. (2008). Comparison of Radiologist and CAD Performance in the Detection of CT-confirmed Subtle Pulmonary Nodules on Digital Chest Radiographs. *Investigative Radiology, 43, No. 6*(June 2008), 343-348.

Brenner, M. W., Holsti, L. R., & Perttala, Y. (1967). The Study by Graphical Analysis of the Growth of Human Tumours and Metastases of the Lung. *The British Journal Of Cancer, XXI*.

Bücheler, E., Lackner, K.-J., & Thelen, M. (2005). *Einführung in die Radiologie.* (Vol. 11). Stuttgart: Thieme

Bush, I. (2016). Lung Nodule Detection and Classification. Retrieved from http://cs231n.stanford.edu/reports/2016/pdfs/313_Report.pdf (07.04.19)

Buxmann, P., & Schmidt, H. (2019). *Künstliche Intelligenz*: Springer Gabler The Cancer Imaging Archive. (2021). Retrieved fromhttps://www.cancerimagingarchive.net/about-the-cancer-imaging-archive-tcia/ (27.12.21)

Castro, D., & New, J. (2016). The Promise of Artificial Intelligence. *Center for Data Innovation*.

Chen, H. (2019). Machine Learning Approaches for IC Manufacturing Yield Enhancement. Retrieved from https://www.researchgate.net/figure/Receiver-Operating-Characteristic-ROC-curves-and-the-area-under-ROC-curve-or-AUC_fig3_331797273 (21.04.21)

Coccia, M. (2020). Deep learning technology for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence. *Technology in Society, 60*. doi:10.1016/j.techsoc.2019.101198

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation, 132 No. 20,* 11.

Deutsches Institut für Medizinische Dokumentation und Information. (2018). ICD-10-GM. Retrieved from https://www.dimdi.de/dynamic/de/klassifikationen/icd/icd-10-gm/ (18.12.20)

Diederich, S., Wormanns, D., & Heindel, W. (2001). Radiologisches Screening des Bronchialkarzinoms: Aktueller Standund zukünftige Perspektiven. *Fortschr Röntgenstr, 173*, 873-882.

Dreher, M., Berthold, J., Nilius, G., Woehrle, H., & Rembert, K. (2019). Stand der digitalen Medizin in der Pneumologie. *Dtsch Med Wochenschr, 144*, 6.

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Pineros, M., . . . Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer, 144*(8), 1941-1953. doi:10.1002/ijc.31937

Firmino, M., Angelo, G., Morais, H., Dantas, M. R., & Valentim, R. (2016). Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. *Biomed Eng Online, 15*, 2. doi:10.1186/s12938-015-0120-7

Flatworldsolutions. (2020). Top 10 Application of Machine Learning in Healthcare Retrieved from https://www.flatworldsolutions.com/healthcare/articles/top-10-applications-of-machine-learning-in-healthcare.php (23.01.21)

Gasparini, S., Ferretti, M., Secchi, E. B., Baldelli, S., Zuccatosta, L., & Gusella, P. (1995). Integration of Transbronchial and Percutaneous Approach in the Diagnosis of Peripheral Pulmonary Nodules or Masses. *Chest Clinical Investigation, 108*, 7.

Gebhardt, G. (2017). Sectra: Integrierte Befundung im PACS. Retrieved from https://www.radiologiemagazin.de/sectra-integrierte-befundung-im-pacs (06.12.20)

Geddes, D. M. (1979). The Natural History of Lung Cancer: A Review Based on the Rates of Tumour Growth.

Giger, M. L. (2018). Machine Learning in Medical Imaging. *J Am Coll Radiol, 15*(3 Pt B), 512-520. doi:10.1016/j.jacr.2017.12.028

Giger, M. L., Karssemeijer, N., & Schnabel, J. A. (2013). Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annu Rev Biomed Eng, 15*, 327-357. doi:10.1146/annurev-bioeng-071812-152416

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. 21. Retrieved from https://arxiv.org/pdf/1311.2524.pdf (29.01.21)

Goertzel, B., & Pennachin, C. (2007). Artifical General Intelligence.

Good, C. A., & Wilson, T. W. (1958). The solitary circumscribed pulmonary nodule. *166*.

Goram, M. (2018a). Künstliche Intelligenz. *Datenbanken verstehen.* Retrieved from http://www.datenbanken-verstehen.de/lexikon/kuenstliche-intelligenz/ (14.12.20)

Goram, M. (2018b). Machine Learning Retrieved from http://www.datenbanken-verstehen.de/lexikon/machine-learning/ (14.12.20)

Hanser, H., & Scholtyssek, C. (2000). symbolische Wissensrepresentation. *Lexikon der Neurowissenschaft.* Retrieved from https://www.spektrum.de/lexikon/neurowissenschaft/symbolische-wissensrepraesentation/12576 (17.12.20)

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Computing Research Repository abs/1512.03385 (2015)*.

Hecker, E., & Ukena, D. (2004). Isolierter Lungenrundherd. *Der Pneumologe, 1*(2), 113-124. doi:10.1007/s10405-004-0014-z

Heyer, C. (2007). Von vielen Ärzten unterschätzt. *Deutsches Ärzteblatt, 30*, A 2145.

Hilgers, R.-D., Heussen, N., & Stanzel, S. (2019). *Lexikon der Medizinischen Laboratoriumsdiagnostik* (3 ed.).

Hoffmann, H., & Dienemann, H. (2000). Der pulmonale Rundherd: Prinzipien der Diagnostik. *Deutsches Ärzteblatt 97*.

Howlader, N., Noone, A., Krapcho, M., J., G., D., M., Altekruse, S., . . . Feuer, E. (2014). Cancer of the Lung and Bronchus (Invasive). In. Bethesda, MD: National Cancer Institute.

Johner, C. (2018, 2015, 27.03.). Verifizierung Und Validierung: Unterschied & Definitionen. Retrieved from https://www.johner-institut.de/blog/iec-62304-medizinische-software/verifizierung-und-validierung-von-medizinprodukten/ (08.09.19)

Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., & Hajirasouliha, I. (2017). Deep Convolutional Neural Networks Enable Discrimination of

Heterogeneous Digital Pathology Images. *EBioMedicine, 27*, 12.

Krüger-Brand, H. E. (2006). PACS – Picture Archiving and Communication System: Abschied von der „Bildertüte“. *Dtsch Arztebl International, 103*(28-29), 1949-. Retrieved from https://www.aerzteblatt.de/int/article.asp?id=52129 (06.12.20)

Lillington, G. A. (1990). Decision Analysis for Management of Solitary Pulmonary Nodules. *Mayo Clin Proc*, 4. doi:https://doi.org/10.1016/S0025-6196(12)65167-2 (21.07.19)

Lillington, G. A. (1991). Management of Solitary Pulmonary Nodules.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2018). Focal Loss for Dense Object Detection. 10. Retrieved from https://arxiv.org/pdf/1708.02002.pdf (14.08.20)

Litzel, S. L. N. (2016). Was ist Machine Learning? Retrieved from https://www.bigdata-insider.de/was-ist-machine-learning-a-592092/ (19.10.20)

MacMahon, H., Naidich, D. P., Goo, J. M., Lee, K. S., Leung, A. N. C., Mayo, J. R., . . . Bankier, A. A. (2017). Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology, 284*, 15. Retrieved from file:///C:/Users/marie/Downloads/radiol.2017161659.pdf https://pubs.rsna.org/doi/full/10.1148/radiol.2017161659 (08.09.21)

Magoulas, G. D., & Prentza, A. (2001). Machine Learning in Medical Applications. In V. K. G. Paliouras, and C.D. Spyropoulos (Ed.), (pp. 300-307): Springer-Verlag.

Mansmann, U. (2011). Biostatistische Methoden - Statistische Inferenz bei ROC Kurven In.

Marten, K., Engelke, C., Seyfarth, T., Grillhosl, A., Obenauer, S., & Rummeny, E. J. (2005). Computer-aided detection of pulmonary nodules: influence of nodule characteristics on detection performance. *Clin Radiol, 60*(2), 196-206. doi:10.1016/j.crad.2004.05.014

Mildenberger, P. D. m. P. (2011) *»Streaming-Lösungen werden das Client-Server-Konzept ablösen«*. Healthcare IT Branchenführer.

Minto, L. (2009). What is the F2 score in machine learning? Retrieved from https://www.quora.com/What-is-the-F2-score-in-machine-learning (19.10.20)

Mintz, Y., & Brodie, R. . (2019). Introduction to artificial intelligence in medicine. doi:10.1080/13645706.2019.1575882

Moeser, J. (2017). Starke KI, Schwache KI – Was kann Künstliche Intelligenz? Retrieved from https://jaai.de/starke-ki-schwache-ki-was-kann-kuenstliche-intelligenz-261/ (23.10.20)

Monnier-Cholley, L., MacMahon, H., Katsuragawa, S., Morishita, J., Ishida, T., & Doi, K. (1998). Computer-aided diagnosis for detection of interstitial opacitieson chest radiographs. *American Journal of Roentgenology, 171*, 6. doi:10.2214/ajr.171.6.9843307

Nakamura, K., Yoshida, H., Engelmann, R., MacMahon, H., Katsuragawa, S., Ishida, K., . . . Doi, K. (2000). Computerized Analysis of the Likelihood of Malignancy in Solitary Pulmonary Nodules with Use of Artificial Neural Networks. *Radiology, 2014*, 8. doi:10.1148/radiology.214.3.r00mr22823

Nam, J. G., Park, S., Hwang, E. J., Lee, J. H., Jin, K. N., Lim, K. Y., . . . Park, C. M. (2019). Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology, 290*(1), 218-228. doi:10.1148/radiol.2018180237

Narkhede, S. (2019). AUC verstehen - ROC-Kurve - Maschinelles Lernen - 2019. Retrieved from https://sciencewal.com/53654-understanding-auc-roc-curve-68b2303cc9c5-20 (24.03.21)

Ngoya, P. S., Muhogora, W. E., & Pitcher, R. D. (2016a). Defining the diagnostic divide: an analysis of registered radiological equipment resources in a low-income African country. *Pan African Medical Journal, 25*, 99. doi:10.11604/pamj.2016.25.99.9736

Ngoya, P. S., Muhogora, W. E., & Pitcher, R. D. (2016b). Defining the diagnostic divide: an analysis of registered radiological equipment resources in a low-income African country. *Pan Afr Med J, 25*, 99. doi:10.11604/pamj.2016.25.99.9736

Ost, D., Fein, A. M., & Feinsilver, S. H. (2003). The Solitary Pulmonary Nodule. *The New England Journal Of Medecine*.

Pape, W. (1914). Handwörterbuch der griechischen Sprache. Bd. 1: A-K. Retrieved from https://de.wikipedia.org/wiki/Diagnose#cite_note-1 (05.06.19)

Peuchot, M., & Libshitz, H. I. (1987). Pulmonary Metastatic Disease: Radiologic-Surgical-Correlation. *Thoracic Radiology, 164*, 4.

Preisner, S. (2011). DICOM, PACS, RIS & Teleradiologie. Retrieved from https://www.mta-r.de/blog/dicom-pacs-ris-teleradiologie/ (07.12.20)

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., . . . Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med, 15*(11), e1002686. doi:10.1371/journal.pmed.1002686

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. 8.

Rubin, G. D., Lyo, J. K., Paik, D. S., Sherbondy, A. J., Chow, L. C., Leung, A. N., . . . Napel, S. (2005). Pulmonary Nodules on Multi–Detector Row CT Scans: Performance Comparison of Radiologists and Computer-aided Detection. *Radiology, 234*, 274-283.

Rubins, J. B., & Bloomfield Rubins, H. (1996). Temporal Trends in the Prevalence of Malignancy in Resected Solitary Pulmonary Lesions. *Chest Clinical Investigations, 109*, 4. doi:10.1378/chest.109.1.100

Scherk, J., Pöchhacker-Tröscher, M. G., & Wagner, K. (2017). Künstliche Intelligenz - Artificial Intelligence.

Schmette, P. (2018). *Lung Cancer Segmentation in Generated Chest X-ray and Darkfield Images using Neural Networks.*

Schmitt, M. (2019). Künstliche Intelligenz in der Medizin. Retrieved from https://www.datarevenue.com/de-blog/kuenstliche-intelligenz-in-der-medizin (20.10.20)

Schober, S. A. (2019). *Deep Learning CNN Methods for Lung Tumor Screening and Detection in Conventional Chest Radiographs.* (Master Thesis). (April 2019)

Schultheiss, M., Schober, S. A., Lodde, M., Bodden, J., Aichele, J., Müller-Leisse, C., . . . Pfeiffer, D. (2020). A Robust Convolutional Neural Network for Lung Nodule Detection in the Presence of Foreign Bodies.

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences, 3*, 417-457.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K., . . . Doi, K. (2000a). Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule

Receiver Operating Characteristic Analysis of Radiologists' Detection of Pulmonary Nodules. *American Journal of Roentgenology, 174*, 4.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K., . . . Doi, K. (2000b). Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule: Receiver Operating Characteristic Analysis of Radiologists' Detection of Pulmonary Nodules. *American Journal of Roentgenology, 174 (1)*, 71-74. doi:doi:10.2214/ajr.174.1.1740071

Tan, B. B., Flaherty, K. R., Kazerooni, E. A., & Iannettoni, M. D. (2003). The Solitary Pulmonary Nodule. doi:10.1378/chest.123.1_suppl.89s

Toomes, H., Delphendahl, A., Manke, H. G., & Vogt-Moykopf, I. (1983). The coin lesion of the lung. A review of 955 resected coin lesions. *Cancer, 51*(3), 534-537.

Trabold, D. (2021). Retrieved from https://machinelearning-blog.de/grundlagen/welche-arten-von-maschinellem-lernen-gibt-es/ (28.10.20)

Tuddenham, W. (1984). Glossary of terms for thoracic radiology: recommendations of the Nomenclature Committee of the Fleischner Society. *143*. Retrieved from https://www.ajronline.org/doi/pdf/10.2214/ajr.143.3.509 (08.09.21)

Turing, A. M. (2021). *Computing Machinery and Intelligence / Können Maschinen denken?* : Reclam.

Tutanch. (2016). Was ist Machine Learning? Retrieved from https://www.bigdata-insider.de/was-ist-machine-learning-a-592092/ (27.10.20)

Van Ginneken, B., Schaefer-Prokop, C. M., & Prokop, M. (2011). Computer-aided Diagnosis:How to Move from the Laboratory to the Clinic *Radiology, 261*, 14.

Van Ginneken, B., Ter Haar Romeny, B. M., & Viergever, M. A. (2001). Computer-Aided Diagnosis in Chest Radiography: A Survey. *IEEE Transactions on Medical Imaging, 20*. doi:10.1109/42.974918

Vehmeyer, S., Scherer, Oberbeck. (2014). Künstliche Intelligenz. Retrieved from https://www.slideserve.com/ayanna-russell/k-nstliche-intelligenz (29.10.20)

Welling, R. D., Azene, E. M., Kalia, V., Pongpirul, K., A., S., Sydnor, R., . . . Mollura, D. J. (2011). *White Paper Report of the 2010 RAD-AID Conference on International Radiology for Developing Countries: Identifying Sustainable Strategies for Imaging Services in the Developing World*. Retrieved from

Wichert, A. (2018, 2010). Künstliche Intelligenz. Retrieved from https://www.spektrum.de/lexikon/neurowissenschaft/kuenstliche-intelligenz/6810 (28.10.20)

Yokota, H., Goto, M., Bamba, C., Kiba, M., & Yamada, K. (2017). Reading efficiency can be improved by minor modification of assigned duties; a pilot study on a small team of general radiologists. *Jpn J Radiol, 35*(5), 262-268. doi:10.1007/s11604-017-0629-8

Yu, K.-H., Zhang, C., Berry, G., Altman, R. B., Ré, C., Rubin, D. L., & Snyder, M. (2016). Predicting nonsmall cell lung cancer prognosis by fully automated microscopic pathology doi:10.1038/ncomms12474

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine, 15*(11), e1002683-e1002683. doi:10.1371/journal.pmed.1002683

Zetkin, M. (1980). *Wörterbuch der Medizin*. München.

VI Danksagung

An dieser Stelle möchte ich allen beteiligten Personen danken, die mich bei der Anfertigung meiner Doktorarbeit unterstützt haben.

Mein besonderer Dank gilt Prof. Dr. Daniela Pfeiffer, meiner Doktormutter, die mir dieses interessante Thema überlassen hat. Ich bedanke mich für die in jeder Hinsicht ausgezeichnete Betreuung und die enorme Unterstützung bei der Umsetzung der gesamten Arbeit. Vielen Dank für all die Antworten auf meine unzähligen Fragen und das umfangreiche Korrekturlesen.

Zudem möchte ich Manuel Schultheiß für die wissenschaftliche Betreuung, Hilfsbereitschaft, als auch seinen unerschöpflichen Rat und die mühevolle Arbeit des Korrekturlesens danken.

Außerdem danke ich Sebastian Schober, der eine tolle Vorarbeit geleistet hat und mir mit Antworten während der Bearbeitung meiner Doktorarbeit zur Seite stand.

Darüber hinaus möchte ich insbesondere Julius Amecke-Mönnighoff Danke sagen, der durch seine wertvollen Anregungen und endlose Geduld, als auch seine uneingeschränkte Motivation maßgeblich zum Gelingen dieser Arbeit beigetragen hat.

Zudem danke ich Dr. Dietrich Sonntag für seine Hilfsbereitschaft und kritische Auseinandersetzung mit meinem Themenkomplex.

Bei meiner Familie und meinen Freunden möchte ich mich herzlich für all das Verständnis und die Unterstützung während meines Studiums und der Doktorarbeit bedanken.

Abschließend bedanke ich mich bei dem gesamten Lehrstuhlteam, sowie bei der Direktion der radiologischen Klinik des Klinikums rechts der Isar, für die Ermöglichung dieser Arbeit und für die Unterstützung bei dessen Umsetzung.