

Aus dem Institut für Diagnostische und Interventionelle Radiologie
der Universität zu Köln
Direktor: Universitätsprofessor Dr. med. D. Maintz

Radiomics-basierte Differenzierung gutartiger Lungenrundherde von Lungenmetastasen

Inaugural-Dissertation zur Erlangung der ärztlichen Doktorwürde
der Medizinischen Fakultät
der Universität zu Köln

vorgelegt von
Thore Struck
aus Hamburg

promoviert am 15. September 2026

Gedruckt mit Genehmigung der Medizinischen Fakultät der Universität zu Köln
2026

Dekanin/Dekan: Universitätsprofessor Dr. med. G. R. Fink

1. Gutachter: Universitätsprofessor Dr. med. T. Persigehl
2. Gutachterin: Privatdozentin Dr. med. L. Nogová

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Dissertationsschrift ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes habe ich Unterstützungsleistungen von folgenden Personen erhalten:

Herrn Univ.-Prof. Dr. Thorsten Persigehl
Herrn Priv.-Doz. Dr. med. Simon Lennartz
Frau Dr. Mirjam Schöneck
Herrn Priv.-Doz. Dr. med. Daniel Zopfs
Frau Alina Mager

Weitere Personen waren an der Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe einer Promotionsberaterin/eines Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertationsschrift stehen.

Die Dissertationsschrift wurde von mir bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die Abfrage der radiologischen Datenbank des Instituts, die potenzielle Proband*innen identifizierte, und die anschließende Zusammenfassung in Listenform wurde durch Herrn Priv.-Doz. Dr. med. Simon Lennartz durchgeführt. Diese arbeitete ich nun nach Einweisung durch Herrn Priv.-Doz. Dr. med. Simon Lennartz, Herrn Priv.-Doz. Dr. med. Daniel Zopfs und Frau Alina Mager Fall für Fall eigenständig durch, indem ich das radiologische Bildmaterial im PACS, die Krankengeschichte und eventuelle histologische Befunde in ORBIS sichtete, um zu überprüfen, ob die Einschlusskriterien erfüllt wurden. Anschließend übertrug ich eigenständig die entsprechenden Bildserien mittels der Software DICOM Query ins Programm mint Lesion research und führte die Segmentierungen der Läsionen durch. Diesbezüglich erfolgte eine Einweisung durch Herrn Priv.-Doz. Dr. med. Simon Lennartz und Herrn Priv.-Doz. Dr. med. Daniel Zopfs. Die Extraktion der Radiomics-Features mittels PyRadiomics 3.0.1 wurde nach Abschluss der Segmentierungen durch Dr. Mirjam Schöneck durchgeführt. Die maschinellen Lernprozesse zur Erstellung der Klassifikationsmodelle und der Erfassung der Interreader-Variabilität wurden ebenfalls durch Dr. Schöneck mittels Python 3.8.13 samt der Packages pandas, NumPy, scikit-learn, XGBoost und Precision-Medicine-Toolkit programmiert und die Ergebnisse durch sie tabellarisiert und graphisch dargestellt.

Erklärung zur guten wissenschaftlichen Praxis:

Ich erkläre hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten (Amtliche Mitteilung der Universität zu

Köln AM 132/2020) der Universität zu Köln gelesen habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen.

Köln, den 29.03.2026

Unterschrift:

Danksagung

Ich bedanke mich bei Herrn Univ.-Prof. Dr. Thorsten Persigehl für die Möglichkeit, meine Promotion am Institut für diagnostische und interventionelle Radiologie durchführen zu können und, für das in mich gesetzte Vertrauen und für die Bereitstellung des Themas.

Herrn Priv.-Doz. Dr. med. Simon Lennartz möchte ich für die langjährige Betreuung danken. Er stand mir von Beginn des Projektes an stets sehr bemüht als wertvoller Ansprechpartner zur Seite.

Dr. Mirjam Schöneck war durch ihre Arbeit im Bereich Data Science und Machine Learning ein unentbehrlicher Teil des Projekts und begegnete mir stets hilfsbereit und unterstützend.

Priv.-Doz. Dr. med. Daniel Zopfs und Frau Alina Mager waren insbesondere zu Beginn des Projekts von großer Hilfe, um mich mit dem korrekten Arbeiten vertraut zu machen.

Außerdem möchte ich mich bei all den Patient*innen bedanken, die ihre Untersuchungen der Forschung zur Verfügung gestellt haben.

Abschließend gilt meiner Familie und meinen Freunden ein großer Dank für ihren Beistand und ihre Unterstützung, insbesondere während des Verfassens dieser Dissertationsschrift.

Diese Dissertation ist meiner Familie gewidmet.

Inhaltsverzeichnis

ABKÜRZUNGSVERZEICHNIS	9
1. ZUSAMMENFASSUNG	10
2. EINLEITUNG	12
2.1. Lungenrundherde: Allgemeines und Epidemiologie	12
2.2. Differenzierung von Lungenrundherden	14
2.2.1. Lungenrundherd-Diagnostik im Kontext des Lungenkarzinoms	15
2.2.2. Differenzierung von Lungenrundherden bei onkologischen Patient*innen	20
2.3. Radiomics	26
2.3.1. Hintergrund	26
2.3.2. Künstliche Intelligenz, maschinelles Lernen und Deep Learning	27
2.3.3. Allgemeiner Arbeitsablauf	29
2.3.4. Limitationen	31
2.4. Anwendung von Radiomics	33
2.4.1. Anwendungen in der Onkologie	33
2.4.2. Radiomics zur Klassifikation unklarer Lungenrundherde	35
2.5. Fragestellungen und Ziel der Arbeit	36
3. MATERIAL UND METHODEN	38
3.1. Kohorte	38
3.1.1. Einschlusskriterien der Proband*innen	38
3.1.2. Einschlusskriterien der Lungenrundherde	38
3.2. Gewinnung und Rekonstruktion des Bildmaterials	41
3.3. Extraktion der Radiomics-Features aus Bilddaten	42
3.3.1. Semiautomatisierte Segmentierung der Lungenrundherde	42
3.3.2. Interreader-Variabilitäts-Analyse	42
3.3.3. Extraktion der Radiomics-Features	42
3.4. Statistische Analyse und Modellentwicklung	43
3.4.1. Entwicklung der Klassifikationsmodelle	46
3.4.2. Analyse von Subgruppen	47

3.5.	Evaluation der methodischen Qualität der Studie	47
4.	ERGEBNISSE	48
4.1.	Kohorte	48
4.2.	Interreader-Variabilität	49
4.3.	Entwicklung der Klassifikationsmodelle	50
4.4.	Klassifikationsergebnisse	51
4.4.1.	Gesamte Kohorte	51
4.4.2.	Rekonstruktionen	53
4.4.3.	Feature-Subgruppen	53
4.5.	Evaluation der methodischen Qualität der Studie	53
5.	DISKUSSION	54
5.1.	Studienkohorte	54
5.2.	Inter- und Intrareader-Variabilität	57
5.3.	Klassifikationsergebnisse	58
5.3.1.	Gesamte Kohorte	58
5.3.2.	Rekonstruktionen	62
5.3.3.	Feature-Subgruppen	63
5.4.	Feature-Selektion	65
5.5.	Limitationen	66
5.6.	Schlussfolgerungen	68
6.	LITERATURVERZEICHNIS	70
7.	ANHANG	80
7.1.	Abbildungsverzeichnis	80
7.2.	Tabellenverzeichnis	80
7.3.	Radiomics Quality Score	80

Abkürzungsverzeichnis

CT	Computertomographie
PET	Positronen-Emissions-Tomographie
MRT	Magnetresonanztomographie
DECT	Dual-Energy-Computertomographie
LDCT	Niedrigdosis-Computertomographie
VDT	Volumenverdopplungszeit
SUV	Standard-Aufnahmewert
18F-FDG	¹⁸ F-Fluordesoxyglucose
ELCAP	Early Lung Cancer Action Project
I-ELCAP	International Early Lung Cancer Action Project
NLST	National Lung Screening Trial
NELSON	Nederlands-Leuvens Longkanker Screenings Onderzoek
VATS	Videoassistierte thorakoskopische Chirurgie
LNT	Linear-No-Threshold-Modell
USA	Vereinigte Staaten von Amerika
ACCP	American College of Chest Physicians
KI	künstliche Intelligenz
ML	Maschinelles Lernen
KNN	künstliches neuronales Netz
IBSI	Image Biomarker Standardization Initiative
ROI	region of interest
IMR	iterative model reconstruction/iterative Modellrekonstruktion
AUC	Area Under the Curve
AUROC	Area Under the ROC Curve
ROC	Receiver operating characteristic
CCC	Konkordanz-Korrelationskoeffizient
RQS	Radiomics Quality Score
CEA	Carcinoembryonales Antigen
EGFR	Epidermal Growth Factor Receptor
KRAS	Kirsten Rat Sarcoma
BRAF	B-rapidly accelerated fibrosarcoma

1. Zusammenfassung

Hintergrund

Inzidentelle Lungenrundherde in Staging-Untersuchungen onkologischer Patienten stellen eine diagnostische Herausforderung in der Radiologie dar. Zu deren näherer Einordnung stehen zum einen engmaschige Follow-Up-Untersuchungen mit Beurteilung des Wachstumsverhaltens, zum anderen invasive Gewebeanalysen durch Biopsien zur Verfügung. CT-Verlaufskontrollen sind jedoch mit erhöhter Strahlenbelastung, Biopsien mit gewissen Komplikationsrisiken verbunden. Somit böten Differenzierungsmodelle basierend auf quantitativen Bildanalysetechniken (Radiomics) eine mögliche risikoarme Alternative. Viele bisherige Studien sind auf die bildbasierte Differenzierung gutartiger Rundherde von primären Lungenkarzinomen fokussiert oder auf Metastasen einzelner Malignome beschränkt. Ziel dieser Arbeit war es daher zu untersuchen, ob mittels Radiomics-basierter Modelle metastatische Rundherde bei Patient*innen mit verschiedenen onkologischen Grunderkrankungen von benignen Befunden unterscheiden werden können. Hierbei wurden Bilddateien verschiedener Computertomographen mit unterschiedlichen Untersuchungsparametern einbezogen.

Materialien und Methoden

Patient*innen, die im Rahmen ihrer onkologischen Staging-CT-Untersuchungen zwischen Mai 2010 und Juli 2020 kontrastmittelverstärkte CT-Untersuchungen des Thorax erhalten hatten, wurden mittels einer Datenbanksuche ermittelt und entsprechend der Einschlusskriterien inkludiert. Zudem wurden Patient*innen einer vorigen Studie mit gleichen Einschlusskriterien in die Kohorte miteinbezogen. Die Rundherde in den kontrastmittelgestützten CT-Bildern wurden semi-automatisiert segmentiert und Radiomics-Features extrahiert. Anschließend wurden im Rahmen von Kreuzvalidierungsverfahren Klassifikationsmodelle entwickelt. Hierfür wurden sowohl automatisierte als auch manuelle Methoden der Selektion robuster Features verwendet. Die Modelle mit der höchsten Klassifikationsleistung wurden zu Ensemble-Klassifikatoren zusammengefasst. Zur Evaluation der finalen Klassifikationsfähigkeit wurden Sensitivität, Spezifität, Area Under the ROC Curve (AUC), F1-Score und Balanced Accuracy sowohl für die ganze Kohorte als auch für die unterschiedlichen Rekonstruktions-Kernel berechnet. Außerdem wurden Subgruppen der Features einzeln untersucht und die Interreader-Variabilität erfasst.

Ergebnisse

Insgesamt 1413 Rundherde von 484 Patient*innen wurden in die Studie eingeschlossen (1037 Metastasen, 376 benigne). Die häufigsten Primärmalignome waren das maligne Melanom

(n=63) und das Kolorektalkarzinom (n=28). Die mittlere AUC des Ensemble-Klassifikators lag bei $0,81 \pm 0,03$ mit einer Sensitivität von $0,72 \pm 0,07$ und einer Spezifität von $0,75 \pm 0,02$ für die korrekte Differenzierung der benignen Rundherde. Der Vergleich verschiedener Rekonstruktions-Filter ergab Variationen der Klassifikationsleistungen, mit einer Balanced Accuracy von $0,76 \pm 0,06$ in der Kohorte des Filters IMR Sharp Plus. Die Untersuchung größenbasierter Features zeigte keinen nennenswerten Einfluss auf die Klassifikationsleistung. Es ergab sich eine geringe Interreader-Variabilität. Im Vergleich der verschiedenen Methoden der Feature-Selektion konnte allenfalls eine gering höhere Klassifikationsleistung unter Anwendung einer manuellen Selektion gegenüber der automatisierten Feature-Selektion beobachtet werden.

Schlussfolgerungen

Die Differenzierung von Lungenmetastasen und gutartigen Lungenrundherden mit Radiomics-basierten Modellen ist in einer großen Kohorte mit hoher Heterogenität bezüglich Grunderkrankungen und CT-Geräten grundsätzlich möglich. Zum jetzigen Zeitpunkt ist die erzielte Genauigkeit für den klinischen Einsatz jedoch noch nicht ausreichend. Die Studie verdeutlicht zudem den Effekt verschiedener Bildrekonstruktionen auf die Radiomics-Features und folglich auf die Genauigkeit der Klassifikation. Führende Limitationen der Studie sind das mehrheitliche Fehlen einer histopathologischen Dignitätsbestimmung der Rundherde, die Heterogenität und das Klassenungleichgewicht des Datensatzes sowie der retrospektive Charakter der Studie und die fehlende externe Validierung.

2. Einleitung

2.1. Lungenrundherde: Allgemeines und Epidemiologie

Die Fleischner Society, eine international anerkannte Fachgesellschaft für Thoraxradiologie, definiert einen Lungenrundherd als eine typischerweise runde, umschriebene Verdichtung im Lungengewebe, die einen maximalen Durchmesser von 30 mm aufweist. Die Abgrenzung zu anderen pulmonalen Verdichtungsstrukturen erfolgt anhand der Größe: Mikronoduli sind bis zu einem Durchmesser von 6 mm beschrieben, während Raumforderungen einen Durchmesser von 30 mm überschreiten.¹

Die bevorzugte bildgebende Modalität zur Beurteilung pulmonaler Rundherde ist die Computertomographie (CT), da sie im Vergleich zur konventionellen Röntgenaufnahme eine deutlich höhere räumliche Auflösung und die Akquisition von Schnittbildern in unterschiedlichen Ebenen innerhalb weniger Sekunden ermöglicht.² Durch technische Anpassungen wie die Variation der Röhrenspannung oder den Einsatz von Rekonstruktionsfiltern lässt sich die Strahlendosis signifikant reduzieren.³ Eine zunehmend etablierte Ergänzung stellt die Kombination aus CT und Positronen-Emissions-Tomographie (PET-CT) zur Messung der Stoffwechselaktivität der Läsionen dar.⁴ Die Magnetresonanztomographie (MRT) findet hingegen in der Diagnostik pulmonaler Rundherde nur eingeschränkt Anwendung. Gründe hierfür sind die längere Untersuchungsdauer mit erhöhter Anfälligkeit für Bewegungsartefakte sowie die geringe Protonen- und Wasserdichte des luftgefüllten Lungengewebes, die die Bildqualität einschränken.⁴

Hinsichtlich ihrer Dichte lassen sich pulmonale Rundherde in solide und subsolide Läsionen unterteilen:⁵ Solide Rundherde zeigen in der CT eine weichteiläquivalente Dichteanhebung, während subsolide Rundherde eine geringere Dichte als Weichteilgewebe aufweisen. Letztere werden in Milchglasrundherde und teilsolide Rundherde differenziert. Milchglasrundherde sind durch eine durchweg so geringe Dichteanhebung gekennzeichnet, dass pulmonale Gefäße und Bronchialwände innerhalb der Läsion sichtbar bleiben. Teilsolide Rundherde enthalten sowohl solide als auch Milchglasanteile.⁵

Lungenrundherden kann eine Vielzahl an Erkrankungen zugrunde liegen (siehe Tabelle 1).⁶⁻⁸

Kategorie	Erkrankung
Maligne Neoplasien	Lungenkarzinome
	Lymphom
	Karzinoid
	Sarkom
	Pulmonale Metastasen
Benigne Neoplasien	Hamartome
	Chondrome

<i>Infektionen</i>	Lipome
	Fibrom
	Respiratorische Papillomatose
	pulmonal metastasiertes Leiomyom
	Mykobakterien
	Mykosen
	Lungenabszess
	Septische Embolie
	Granulom
	Infektionen durch Nocardia spp.
<i>Immunologische Erkrankungen</i>	Zystische Echinokokkose
	Q-Fieber
	Rheumatoide Arthritis
	Kryptogen oder sekundär organisierende Pneumonie
	Noduläre Sarkoidose
	Granulomatose mit Polyangiitis
	Lymphoide Granulomatose
	Nekrotisierende sarkoide Granulomatose
	Bronchialatresie mit Bronchozele
	Bronchogene Zyste
<i>Kongenitale Abnormalitäten</i>	Bronchopulmonale Sequester
	Intrapulmonales Hämatom
<i>Vaskuläre Erkrankungen</i>	Lungeninfarkt
	Arteriovenöse Malformation
	Intrapulmonale Varize
	Lipoide Pneumonie
<i>Weitere</i>	Amyloidose/Amyloidom
	Subpleurale und endoparenchymale Lymphknoten
	Runde Atelektasen
	Inflammatorischer Pseudotumor

Tabelle 1: Differenzialdiagnosen von Lungenrundherden

Eine Metaanalyse von Chen et al. aus dem Jahr 2024, die insgesamt 24 Studien einschloss, zeigte, dass in 27 % aller CT-Untersuchungen der Lunge Rundherde entdeckt wurden. Signifikante Geschlechterunterschiede konnten nicht nachgewiesen werden.⁹ An dieser Stelle muss erwähnt werden, dass epidemiologische Daten zu Lungenrundherden häufig aus Screeningstudien von Risikopatient*innen für Lungenkarzinome und/oder Staginguntersuchungen stammen und daher nur eingeschränkt auf die Allgemeinbevölkerung übertragbar sind. Das et al. verglichen daher CT-Untersuchungen anderer Indikationen, in denen Lungenrundherde nachgewiesen wurden. Dabei zeigte sich eine erhebliche Variabilität

der Nachweisraten mit Werten zwischen 0,3 % und 27,7 %, bei einer mittleren Detektionsrate von 7,72 %.¹⁰

Die o.g. genannten Subtypen kommen unterschiedlich häufig vor. In den Studien von Li et al.¹¹ und McWilliams et al.¹² waren 62 % - 79,8 % der Rundherde solide, 0,9 % - 26 % teilsolide und 9,3 % - 15,8 % Milchglasherde. In der Early Lung Cancer Action Project (ELCAP)-Studie wurden 81 % solide und 17 % subsolide Herde (darunter 64 % Milchglasherde und 36 % teilsolide Herde) entdeckt.¹³ Subsolide Herde wurden aufgrund der geringeren Dichte vor der Einführung der Computertomographie auf Röntgenuntersuchungen des Thorax nur selten nachgewiesen.¹⁴

Angaben zur durchschnittlichen Anzahl an Rundherden variieren ebenfalls. In einer 2017 veröffentlichten Studie konnte gezeigt werden, dass innerhalb einer großen Population von 3392 Patient*innen, die diagnostizierte Lungenrundherde aufwiesen, 51,5 % einen einzelnen Rundherd hatten, 23,6 % zwei Rundherde, 10,4 % drei Rundherde, 5,6 % vier Rundherde und 8,9 % mehr als 4 Rundherde.¹⁵ Eine weitere Studie, die ebenfalls Patient*innen mit erhöhtem Risiko für Lungenkarzinome untersuchte, zeigte dagegen, dass zu Beginn der Studie durchschnittlich $6,2 \pm 4,0$ Herde vorlagen.¹²

Insgesamt zeigt sich eine kontinuierliche Zunahme der Detektionsrate. Gould et al. berichteten für die Vereinigten Staaten von Amerika (USA) einen Anstieg von 3,9 auf 6,6 entdeckte Rundherde pro 1000 Personenjahre zwischen 2006 und 2012, ohne dass ein vergleichbarer Anstieg der Lungenkarzinomdiagnosen zu verzeichnen war, was zudem auf eine vermehrte Detektion benigner Läsionen und Metastasen hinweist.¹⁶

Angesichts dieser Vielzahl potenzieller Differenzialdiagnosen sowie ihrer steigenden Prävalenz, insbesondere in Risikopopulationen für primäre pulmonale Malignome, kommt der Abklärung der Dignität pulmonaler Rundherde eine zentrale klinische Bedeutung zu. Eine präzise und zügige diagnostische Einordnung ist essenziell, um im Falle einer malignen Genese unverzüglich eine adäquate Therapie einleiten zu können, da dies maßgeblich den weiteren Krankheitsverlauf und die Prognose beeinflusst.¹⁷

2.2. Differenzierung von Lungenrundherden

Die vorliegende Studie untersucht die Radiomics-gestützte Differenzierung benigner pulmonaler Rundherde von Lungenmetastasen bei onkologischen Patient*innen. Der überwiegende Teil der bisherigen Forschung sowie die etablierten diagnostischen Leitlinien für pulmonale Rundherde unklarer Dignität sind primär auf die Detektion und Klassifikation von Lungenkarzinomen ausgerichtet ist, während Lungenmetastasen nur unzureichend berücksichtigt werden. Zur Verdeutlichung dieses Ungleichgewichts erfolgt im Folgenden zunächst eine Darstellung des aktuellen Forschungsstandes und der geltenden Leitlinien zur radiologischen Diagnostik pulmonaler Primärmalignome. Anschließend wird dargelegt, ob und

in welchem Umfang diese Konzepte auf die Diagnostik metastatischer Rundherde bei Krebspatient*innen übertragbar sind und welche Limitationen hierbei bestehen.

2.2.1. Lungenrundherd-Diagnostik im Kontext des Lungenkarzinoms

Lungenkarzinome zählen weltweit zu den am häufigsten diagnostizierten malignen Neoplasien.¹⁸ In Deutschland stellt Lungenkrebs die häufigste Krebstodesursache bei Männern und bei Frauen die zweithäufigste Krebstodursache nach dem Mammakarzinom dar.¹⁹ Da Lungenkarzinome häufig erst in fortgeschrittenen Stadien entdeckt werden und aufwändige und kostenintensive Therapien erfordern, wurden in den letzten Jahrzehnten verschiedene Maßnahmen zur Optimierung der Diagnostik ergriffen.²⁰ Eine der zentralen Maßnahmen stellen Screening-Untersuchungen für Risikopopulationen dar, um frühestmögliche Diagnosen zu erreichen. Um inzidentelle Lungenrundherdbefunde genauer einordnen zu können, werden verschiedenste bildmorphologische Charakteristika berücksichtigt. Außerdem wurden zum Zweck der Vereinfachung der Beurteilung und zur Vereinheitlichung diagnostischen Vorgehen Klassifikationsmodelle und Handlungsempfehlungen entwickelt. In den folgenden Abschnitten werden diese Aspekte (Screening, radiologische Kriterien, Risikostratifizierungsmodelle und Empfehlungen von Fachgesellschaften) genauer beleuchtet.

2.2.1.1. Screening für Lungenkarzinome

In den 1970er-Jahren wurden mehrere Screening-Studien durchgeführt, die Patient*innen mit erhöhtem Risiko für Lungenkarzinome mittels konventioneller Thoraxradiographie und Sputumzytologie untersuchten. Diese Untersuchungen konnten jedoch keine signifikante Reduktion der Mortalität nachweisen.^{21,22} Im Rahmen des „Prostate, Lung, Colorectal and Ovarian Trial“ wurden insgesamt 154.901 Proband*innen ohne spezifisches Risikoprofil über einen Zeitraum von vier Jahren jährlich mittels Thoraxröntgen untersucht, ebenfalls ohne Nachweis einer Mortalitätsreduktion.²³

In den 1990er-Jahren wurden vermehrt Studien zur Früherkennung von Lungenkarzinomen mittels Niedrigdosis-Computertomographie (LDCT) initiiert. Die erste dieser Untersuchungen war die prospektive Kohortenstudie „Early Lung Cancer Action Project“ (ELCAP), die im weiteren Verlauf in Kooperation mit europäischen und japanischen Zentren zum „International Early Lung Cancer Action Project“ (I-ELCAP) ausgeweitet wurde.²⁴ Hierbei zeigte sich, dass durch LDCT sowohl mehr Lungenkarzinome detektiert als auch frühere Diagnosen gestellt werden konnten als mit konventioneller Röntgendiagnostik.²⁵ Diese Ergebnisse führten zur Durchführung randomisierter kontrollierter Studien, von denen der „National Lung Screening Trial“ (NLST) in den USA eine Reduktion der lungenkrebsbedingten Mortalität um ≥ 20 % nachweisen konnte²⁶ und der niederländisch-belgische „Nederlands-Leuvens Longkanker

Screenings Onderzoek“-Studie (NELSON) nach 10 Jahren eine Mortalitätsreduktion von 24 % in Männern und 33 % in Frauen berichtete.²⁷ Die vielversprechenden Ergebnisse dieser Studien führte zu einer Einführung des Lungenkrebscreenings in der Vereinigten Staaten von Amerika mittels LDCT für Patient*innen mit erhöhtem Erkrankungsrisiko.²⁰

In Deutschland konnte eine Screenings-Studie vergleichbare Ergebnisse wie der NLST zeigen.²⁸ 2024 verabschiedete das Bundesministerium für Umwelt, Klimaschutz, Naturschutz und nukleare Sicherheit die Lungenkrebs-Früherkennungs-Verordnung, welche die Rahmenbedingungen für künftige Screening-Protokolle darlegte.²⁹ Jährliche LDCT-Screening-Untersuchungen für starke Raucher*innen zwischen 50 und 75 Jahren sollen ab April 2026 zu einer Leistung der gesetzlichen Krankenversicherungen werden.³⁰

Die zunehmende Durchführung von Screeninguntersuchungen sowie die kontinuierliche Weiterentwicklung und Verbreitung der Computertomographie haben in den vergangenen Jahren zu einer vermehrten Detektion unklarer pulmonaler Rundherde geführt.^{7,31} Eine genaue diagnostische Abklärung dieser Rundherde ist dadurch zunehmend ins medizinische Interesse gerückt.

2.2.1.2. Radiologische Kriterien zur Rundherddiagnostik im Kontext des Lungenkarzinoms

Eine Vielzahl radiologischer Kriterien werden bei der Evaluation von unklaren Lungenrundherden berücksichtigt.

Die Größe eines Rundherds ist einer der meistdiskutierten Risikofaktoren und wird in der Regel mithilfe des Durchmessers oder des Volumens erfasst. In der NELSON-Studie zum Lungenkrebs-Screening mit insgesamt 9681 Rundherden wurde keine Krebsrisikoerhöhung in Patient*innen mit Rundherden <5 mm gefunden (Risiko = 0,4 %). Ab einem Durchmesser von 5 mm erhöhte sich das Risiko jedoch signifikant und ab einem Durchmesser von >10 mm lag das Risiko bei über 15 %. Ein ähnlicher Zusammenhang wurde bei der Volumetrie der Rundherde entdeckt: Das Vorhandensein von Rundherden unter 100 mm³ zeigte keine signifikante Erhöhung des Krebsrisikos (0,6 %) im Vergleich zu Patient*innen ohne Rundherde. Zwischen 100 und 300 mm³ stieg das Risiko auf 2,4 % an und Herde >300 mm³ wiesen ein Risiko von 16,9 % auf.³² Henschke et al. zeigten vergleichbare Ergebnisse.²⁴ Beim Vorliegen multipler Rundherde ist der größte von ihnen in 80 - 97 % Fällen maligne.^{12,15} Darüber hinaus konnte gezeigt werden, dass bei benignen Rundherden das Verhältnis des maximalen transversalen und des maximalen longitudinalen Durchmessers ($1,62 \pm 0,57$) mehrheitlich größer ist als bei malignen Rundherden ($1,19 \pm 0,23$) sowie malignen Vorstufen ($1,21 \pm 0,21$).³³

Generell ist ein stetig wachsender Rundherd mit hoher Wahrscheinlichkeit maligner Genese.⁶ Es wird angenommen, dass ein Rundherd, der über zwei Jahre kein Wachstum zeigt, mit hoher Wahrscheinlichkeit benigne ist.¹⁴ Eine sensitive Erfassung des Wachstums eines Rundherds ist die Volumenverdopplungszeit (VDT). Allgemein lässt sich näherungsweise sagen, dass benigne Rundherde meist eine sehr geringe VDT (<30 Tage), in der Regel im Falle von entzündlichen Prozessen, oder eine VDT oberhalb von 400 Tagen im Falle von z.B. Hamartomen oder post-infektiösen Granulomen zeigen. Maligne Rundherde dagegen zeigen regelhaft eine VDT von 30-400 Tagen.⁸

Studienergebnisse bezüglich der Rundherdeanzahl als Risikofaktor für Malignität sind widersprüchlich. McWilliams et al. entdeckten eine lineare Reduktion des Malignitätsrisikos mit steigender Zahl der Rundherde.¹² Heuvelmans et al. zeigten zwar eine Spitze der Lungenkrebswahrscheinlichkeit beim Vorhandsein von vier Rundherden, fanden jedoch keinen signifikanten Zusammenhang zwischen Anzahl der Rundherde und Malignität.¹⁵

Auch die Lokalisation eines Rundherds kann Hinweise darauf liefern, ob er benigner oder maligner Genese ist. Krebserkrankungen der Lunge treten am häufigsten in der rechten Lunge, insbesondere peripher sowie im oberen Lungenlappen, auf.³⁴ Solide Rundherde, die perifissural gelegen sind, weisen in der Regel kein Malignitätsrisiko auf und sind mit hoher Wahrscheinlichkeit Lymphknoten.^{12,35,36} Zudem wurde eine subpleurale Lokalisation ebenfalls als Benignitätshinweis beschrieben.³³

Die Form und Beschaffenheit der Ränder eines Rundherds kann Hinweise auf Benignität oder Malignität geben, jedoch bestehen ausgeprägte Überschneidungen zwischen den beiden Klassen. Benigne Rundherde treten meist rund und glatt berandet oder polygonal auf.^{8,33,37} Malignome der Lunge zeigen häufig irreguläre Wachstumsmuster. Darunter fallen radiäre Ausläufer, sogenannte Spikulierungen, die die Ausbreitung der Malignomzellen ins Interstitium repräsentieren sowie ein gelapptes Auftreten durch inhomogene Wachstumsvorgänge. Beide dieser Charakteristika können jedoch auch bei benignen Rundherde inflammatorischer Genese vorkommen.³⁷

In der Computertomographie werden Abschwächungen von ≥ 200 Hounsfield-Einheiten als Maßstab für Verkalkungen verwendet.³⁸ Es werden eine Vielzahl an Kalzifikationsmustern in pulmonalen Rundherden beschrieben, darunter dicht zentral, diffus, lamelliert, Popcorn, punktiert, dendriform, retikulär³⁸ sowie exzentrisch.⁶ Für Benignität sprechen dichte zentrale, laminare und diffuse Verkalkungen. Diese entsprechen meist Granulomen.³⁹ Popcorn-Kalzifikationen sind pathognomonisch für benigne Hamartome.³⁸ Die anderen Kalzifikationsmuster geben keine klaren Hinweise auf die Dignität eines Rundherds und können daher auch maligner Genese sein.³⁸ Es wurden Verkalkungen in bis zu 10 % aller Lungenkarzinome beschrieben.⁴⁰ Darüber hinaus können pulmonale Malignome im Rahmen oder nach Abschluss einer antineoplastischen Therapie verkalken.⁴¹

Auch die Umgebung eines Rundherds kann einen Hinweis auf dessen Dignität darstellen. Ein Halo, ein milchglasartiger Saum um einen solitären Rundherd⁵, kann neben Infektionserkrankungen auch bei verschiedenen Karzinomen auftreten.⁴² Das Reverse-Halo-Phänomen, ein solider Halo um einen Milchglasherdd, dagegen kommt meist bei infektiösen und entzündlichen Vorgängen vor, z.B. bei kryptogen organisierender Pneumonie.⁴³

Obwohl die Ausbildung von Aushöhlungen, sogenannten Kavitäten, meist als Zeichen für Malignität gesehen wird⁴⁴, können diese Strukturen sowohl in benignen als auch malignen Lungenrundherden vorkommen.⁴⁵ Zur Unterscheidung zwischen beiden Entitäten kann die maximale Dicke der Wand, die die Kavitation umgibt, beurteilt werden.⁴⁶

Das Vorliegen von Fett, insbesondere in Kombination mit den bereits erwähnten Popcorn-Kalzifikationen, ist nahezu pathognomonisch für das Vorliegen von benignen Hamartomen.⁴⁷

Die Aufnahme von ¹⁸F-Fluordesoxyglucose (18F-FDG) ist Maß der Stoffwechselaktivität in der PET.⁷ Diese ist in der Regel bei malignen Neoplasien erhöht und 18F-FDG wird vermehrt aufgenommen.⁸ Traditionell wird ein standardisierter Aufnahmewert (SUV) von >2,5 als PET-positiv und dadurch malignitätsverdächtig gesehen.⁴⁸ Limitierend für die PET-CT-Diagnostik ist die Größe der Rundherde. Ab Durchmessern <1cm, insbesondere <8 mm, sinkt die Sensitivität dieser Modalität und die falsch-negativ-Rate steigt.⁴⁴ Falsch-negative Ergebnisse kommen außerdem bei bestimmten Tumorentitäten vor, falsch-positive Ergebnisse mit hohen SUV-Werten kommen vermehrt im Rahmen von Infektionen, Rheumatoider Arthritis, Sarkoidose und endemischen Mykosen vor.⁴⁴

Subsolide Lungenrundherde sind anspruchsvoller in der Diagnostik. Ihre visuelle Beurteilung ist aufgrund der unscharfen Begrenzungen und dem geringen Kontrast zum umgebenden Lungenparenchym schwierig.⁴⁹ Zum einen sind die volumetrischen Messungen schwierig bei diesen Herden anzuwenden¹², zum anderen sind die genannten Richtwerte für VDT nicht generell auf subsolide Herde anwendbar, da die VDT subsolider Herde oft länger sind als die solider Herde.^{14,50,51} Außerdem erschweren subsolide Rundherde auch weitere Diagnostik-Modalitäten: Sie zeigen eine geringe metabolische Aktivität und sind daher für PET-CT-Untersuchungen zur Dignitätsbestimmung weniger gut geeignet.⁴²

2.2.1.3. Modelle zur Risikostratifizierung für das Vorliegen von Lungenkarzinomen

Über die Jahre wurden verschiedene Modelle entwickelt, um das Malignitätsrisiko unklarer pulmonaler Rundherde abzuschätzen. Das erste dieser Art war das Modell der Mayo Clinic (Rochester, USA) aus dem Jahre 1997. Es berücksichtigt radiologische Charakteristika der Rundherde (Durchmesser, Spikulierungen und Lokalisation im oberen Lungenlappen) sowie klinische Faktoren (Alter, Raucherstatus, Vorgeschichte extrathorakaler Malignome) und konnte in Validierungsstudien eine suffiziente Unterscheidung von benignen und malignen

Lungenrundherden zeigen.⁵² Seitdem sind eine Vielzahl ähnlicher Modelle entwickelt worden, unter anderem das Modell des Department of Veterans Affairs und der „Veterans Affairs Cooperative Study“⁵³, das Modell des Peking University People's Hospital⁵⁴ und das PanCan-Modell oder Brock-Modell, welches auf Basis der Daten der „Pan-Canadian Early Detection of Lung Cancer Study“ entwickelt wurde und die höchste AUC der genannten Modelle erreichen konnte (0,970).¹²

Diesen klassischen, klinisch angewandten Modellen stehen neuere gegenüber, die auf dem Prinzip des Deep Learning (siehe 2.3.2.) basieren. Diese zeigen häufig äquivalente bis bessere Ergebnisse als die traditionellen Modelle, werden bisher jedoch nur vereinzelt klinisch eingesetzt.⁵⁵⁻⁵⁷

2.2.1.4. Empfehlungen und Frameworks zur Rundherd-Diagnostik

Zwar liefern die Bildmorphologie der Rundherde und Risikostratifizierungsmodelle wichtige Hinweise zur Dignitätsbestimmung, jedoch geben sie keine allgemeingültigen Empfehlungen für das weitere diagnostische Vorgehen. Leitlinien verschiedener Fachgesellschaften versuchen, einheitliche Handlungsempfehlungen zum Management zu formulieren, um eine möglichst effektive Diagnostik zu ermöglichen.

Zu den prominentesten Empfehlungen gehören die der Fleischner Society, die zuletzt 2017 aktualisiert wurden.⁵⁸ Für solide Rundherde wird hier zunächst eine grundlegende Einschätzung des Malignitätsrisikos des Rundherds in niedrig oder hoch empfohlen. Diese berücksichtigt viele der unter 2.2.1.1. erwähnten Risikofaktoren. Je nach Malignitätsrisiko, Größe der Rundherde und Vorliegen einzelner oder mehrerer Rundherde, werden verschiedene Empfehlungen für die weitere Diagnostik (Zeitraum für Folgeuntersuchungen, PET-CT, Biopsien) ausgesprochen. Für subsolide Rundherde werden Empfehlungen ebenfalls nach Größe, einzeitigem oder multiplen Auftreten sowie teilsolidem oder milchglasartigem Charakter erstellt.

Die Fleischner-Leitlinien schließen verschiedene Patient*innengruppen aus. Personen <35 Jahren sowie onkologische und immunsupprimierte Patient*innen sind von den Empfehlungen ausgenommen. Bei jungen Patient*innen wird empfohlen, die weitere Diagnostik und eventuelle Behandlung individuell anzupassen und die Strahlenbelastung zu reduzieren.⁵⁸

Neben den Fleischner-Kriterien gibt es weitere angewandte Leitlinien. Das American College of Chest Physicians (ACCP) veröffentlicht Empfehlungen zum Rundherde-Management. Grundlage dieser Leitlinien ist ebenfalls eine Unterteilung einzelner oder multipler Rundherde in solide und subsolide. Erstere werden anhand ihres Durchmessers weiter unterteilt, letztere in teilsolide Rundherde und Milchglasherde. Die Einschätzung des Malignitätsrisikos der soliden Rundherde bestimmt ebenfalls deren Management. Das Risiko wird in sehr gering (<5

%), gering/moderat (5–65 %) und hoch (>65 %) eingeteilt. Hierzu werden entweder die Nutzung etablierter Risikostratifizierungsmodelle (siehe 2.2.3.) oder klinisches Urteilsvermögen empfohlen. Je nach Größe, Dichte und Malignitätsrisiko werden Empfehlungen für das weitere diagnostische Vorgehen ausgesprochen, jedoch wird auch die Wichtigkeit des Shared Decision Making betont.⁴⁴

Die British Thoracic Society (BTS) veröffentlichte zuletzt 2015 ihre Leitlinien. Anders als die anderen erwähnten Guidelines präsentiert die BTS ihre Empfehlungen u.a. in Form von vier Algorithmen. Diese befassen sich mit folgenden Themen: initiale Diagnostik und Evaluation solider und nicht-kalzifizierter Herde, Überwachung solider Rundherde, Management subsolider Rundherde und Behandlung von Rundherden.⁵⁹

Das American College of Radiology entwickelte das sogenannte *Lung Imaging Reporting and Data System* (Lung-RADS), das zuletzt 2022 aktualisiert wurde⁶⁰. Diese Leitlinien dienen dazu, die Dokumentation und die weitere Diagnostik von unklaren Rundherden, die im Rahmen von LDCT-Screening-Programmen entdeckt wurden, zu standardisieren. Die Rundherde erhalten entsprechend ihres Malignitätsrisikos einen numerischen Score und je nach Subtyp und Größe werden unterschiedliche Management-Empfehlungen ausgesprochen.⁶¹

Darüber hinaus veröffentlicht das National Comprehensive Cancer Network (NCCN) der USA regelmäßig Leitlinien zum Lungenkrebs-Screening. Diese enthalten zudem Empfehlungen zum Vorgehen bei der Detektion unklarer Rundherde.⁶²

2.2.2. Differenzierung von Lungenrundherden bei onkologischen Patient*innen

Wie im vorherigen Abschnitt ausführlich dargestellt, ist die radiologische Diagnostik von Lungenkarzinomen umfassend untersucht und es existieren mehrere etablierte Empfehlungen, die das diagnostische Vorgehen bei unklaren Lungenrundherden insbesondere in Risikokollektiven für das Vorliegen von Lungenkarzinomen vorgeben. Allerdings sind diese Empfehlungen primär fokussiert auf die Detektion von Lungenkarzinomen.

Bereits diagnostizierte Krebspatient*innen verschiedener onkologischer Grunderkrankungen stellen im Kontext der Rundherddiagnostik ein besonderes Personenkollektiv dar. Im Rahmen von (Re-)Staging-Untersuchungen erfolgen bei dieser Gruppe regelmäßige Computertomographien, bei denen inzidentell unklare Lungenrundherde detektiert werden können. Capracica et al. führten eine retrospektive Analyse von 228 Rundherden bei Patient*innen mit extrapulmonalen Malignomen durch. Die histologische Untersuchung ergab, dass nahezu zwei Drittel der Herde Metastasen der Primärtumore darstellten. Die Gruppe der nicht-metastatischen Herde setzte zu mehrheitlich aus neu diagnostizierten Lungenmalignomen (60 %) und pneumonischen Herden (10 %) zusammen.⁶³ Der korrekten Klassifikation solcher Rundherde kommt eine besondere klinische Bedeutung zu, da das

Vorliegen einer metastasierten Krebserkrankung in der Regel mit einer deutlich ungünstigeren Prognose assoziiert ist und nicht selten zu einer Therapiezieländerung führt.⁶⁴

In den folgenden Abschnitten wird die bestehende Datenlage für die Rundherddiagnostik bei Krebspatient*innen dargestellt sowie Nachteile und potenzielle Risiken erläutert, die sich aus den derzeitigen diagnostischen Vorgehensweisen ergeben.

2.2.2.1. Beurteilung von unklaren Lungenrundherden in Krebspatient*innen

Für die Beurteilung eines inzidentellen Rundherdes bei onkologischen Patient*innen ist zunächst die zugrunde liegende Primärerkrankung von Bedeutung, da das pulmonale Metastasierungspotenzial erheblich variiert. Die häufigsten Primärtumoren mit pulmonaler Filialisierung sind Lungenkarzinome, Kolorektalkarzinome, Nierenzellkarzinome, Pankreaskarzinome und Mammakarzinome.⁶⁵

Im Vergleich zur Diagnostik primärer Lungenkarzinome ist die Evidenzlage zur radiologischen Klassifikation pulmonaler Metastasen deutlich schwächer. Radiomorphologische Kriterien, die eine verlässliche Differenzierung zugunsten einer metastatischen Genese erlauben, sind bislang kaum beschrieben. Ein solides, rundes oder polygonales Erscheinungsbild wird bei Metastasen zwar häufiger beschrieben als bei Karzinomen^{33,37,66,67}, kommt jedoch ebenso und insbesondere bei benignen Rundherden vor und besitzt daher keinen ausreichenden differenzialdiagnostischen Wert. Das Auftreten multipler Rundherde im Kontext einer aktiven malignen Grunderkrankung ist insgesamt als hochverdächtig für eine metastatische Progression zu werten⁶³, kann jedoch auch bei nicht-malignen Erkrankungen wie der Sarkoidose beobachtet werden.⁶⁸

Kleine Rundherde werden in der Regel als höchstwahrscheinlich benigne eingestuft, jedoch bezieht sich diese Information auf Personen ohne bekannte maligne Grunderkrankung.³²

Autopsiestudien verstorbener Krebspatient*innen zeigten, dass Lungenläsionen mit einem Durchmesser zwischen 1 und 5 mm in nahezu 60 % der Fälle metastatisch waren.⁶⁹ Dies legt nahe, dass selbst kleinste Läsionen onkologischer Patient*innen ein signifikantes Malignitätspotential aufweisen und einer zügigen Abklärung bedürfen. Die PET-CT weist bei kleinen Läsionen, wie unter 2.2.1.1. beschrieben, nur eine eingeschränkte diagnostische Aussagekraft auf, da bei Durchmessern <1 cm die falsch-negativ-Rate steigt.⁴⁴

Kavitationen werden zwar seltener bei Metastasen als bei Primärmalignomen beschrieben⁷⁰, kommen jedoch durchaus vor.⁶⁷ Es gibt jedoch auch zahlreiche benigne Differenzialdiagnosen (siehe 2.2.1.1.).⁴⁵ Kalzifikationen treten u.a. bei Metastasen von Sarkomen und Adenokarzinomen vor⁶⁷, pathognomische Kalzifikationsmuster für Metastasen werden nicht beschrieben.⁴⁵

Ein wesentlicher diagnostischer Parameter zur Beurteilung der Genese eines unklaren pulmonalen Rundherdes ist das Wachstumsverhalten. Eine rasche Größenzunahme stellt grundsätzlich einen Malignitätshinweis dar.⁶ Ähnlich wie bei den verschiedenen Subtypen pulmonaler Karzinome, ist das Wachstumsverhalten pulmonaler Metastasen jedoch nicht homogen, sondern zeigt deutliche Unterschiede je nach zugrundeliegendem Primärtumor. So wurde für Sarkome eine Volumendopplungszeit (VDT) von etwa 25 Tagen beschrieben⁷¹, während für Nierenzellkarzinome eine mittlere VDT von 89,4 Tagen ermittelt wurde.⁷² Insgesamt beträgt die VDT pulmonaler Metastasen in der Regel weniger als ein Jahr, wobei insbesondere mesenchymale Malignome erhebliche interindividuelle Variationen aufweisen.⁷³

Explizite Handlungsempfehlungen für unklare Lungenrundherde bei Krebspatient*innen bestehen jedoch nicht: Die Fleischner-Kriterien schließen onkologische Patient*innen explizit aus⁵⁸, Lung-RADS ist ausschließlich für im Rahmen von LDCT-Screening-Programmen identifizierte Rundherde konzipiert.⁶² Die BTS-Guidelines differenzieren nicht zwischen Metastasen und Lungenkarzinomen und sind ebenfalls primär auf Screening-Szenarios ausgerichtet.⁵⁹ ACCP-Guidelines nehmen ebenso keine Differenzierung der Malignome vor.⁴⁴ Darüber hinaus liegen nach Kenntnisstand des Autors keine Modelle zur Einschätzung des Metastasenrisikos vor. Dieser Mangel an spezifischen Vorgaben erschwert die standardisierte Bildbefundung und -beurteilung und führt folglich in der Praxis dazu, dass Empfehlungen für das weitere diagnostische Vorgehen individuell erfolgen müssen, was häufig unter anderem zu engmaschigen Kontrolluntersuchungen und/oder invasiven Probeentnahmen führt.⁶⁴

Aus diesen Gründen stellen Rundherde unklarer Dignität bei onkologischen Patient*innen, insbesondere wenn diese noch sehr klein sind und solitär auftreten, eine besonders schwierige diagnostische Herausforderung dar, die aktuell noch nicht ausreichend erforscht und reguliert ist.

2.2.2.2. Beurteilung von unklaren Lungenrundherden in Krebspatient*innen

Aufgrund der unzureichenden Datenlage zur Diagnostik unklarer Lungenrundherde bei onkologischen Patient*innen, bleibt die Dignität der Läsionen häufig unklar. Wie bereits im vorigen Abschnitt beschrieben (siehe 2.2.2.1.) führt dies in der Regel zu engmaschigen Verlaufsbildgebungen oder Biopsien. Diese sind mit einer erhöhten Belastung an ionisierender Strahlung bzw. den Risiken einer invasiven Probeentnahme verbunden.

Zur definitiven Abklärung unklarer Läsionen steht weiterhin nur die Biopsie mit anschließender histopathologischer Untersuchung zur Verfügung. Diese wird im insbesondere Falle von hochgradig suspekten Rundherden empfohlen, um den Beginn einer eventuellen antineoplastischen Therapie oder eine Anpassung der laufenden Therapie bei Vorliegen einer

Fernmetastasierung nicht weiter zu verzögern.⁵⁸ Verschiedene Verfahren kommen zum Einsatz:

Für nicht-operative Probeentnahmen stehen sowohl CT-gesteuerte perkutane (Stanzbiopsie und Feinnadelaspiration) als auch bronchoskopische Verfahren zur Verfügung.^{74,75} Die Computertomographie ermöglicht eine sehr gute dreidimensionale Planung und Steuerung⁷⁶ und sowohl Stanzbiopsien als auch Feinnadelaspiration zeigen Sensitivitäten und Spezifitäten von $\geq 90\%$.⁷⁷ Bronchoskopische Verfahren zeigten lange deutlich geringere Sensitivitäten und hatten daher einen geringen diagnostischen Stellenwert⁴⁴. Verschiedene technische Weiterentwicklungen konnten die Lokalisation von insbesondere peripher gelegenen Rundherden verbessern und das Verfahren beschleunigen⁷⁵, sodass die Sensitivitäten beider Verfahren für Rundherde zwischen 20 und 30 mm nun ähnlich hoch sind.⁷⁸

Der große Nachteil von Biopsien ist deren invasiver Charakter und das dadurch entstehende Risiko für Komplikationen. Heerink et al. führten 2017 eine Metaanalyse der Risiken der CT-gesteuerten transthorakalen Biopsie durch und zeigten, dass in insgesamt 38,8 % der Stanzbiopsien und 24,0 % der Feinnadelaspirationen Komplikationen auftraten. Davon traten schwerwiegende Komplikationen, namentlich Hämatothorax, Pneumothorax mit Interventionsbedarf, Luftembolie, Stichkanalmetastasen und Tod in 5,7 % der Stanzbiopsien und 4,4 % der Feinnadelaspirationsbiopsien auf.⁷⁴ Wiener et al. zeigten, dass es in 15 % der perkutanen Biopsien zur Entwicklung eines Pneumothorax kam, von denen 6,6 % mittels Thoraxdrainage behandelt werden mussten. Blutungen waren mit 1 % seltener, jedoch benötigten 17,8 % der Betroffenen eine Bluttransfusion. Patient*innen, die Komplikationen erlitten, zeigten außerdem ein erhöhtes Risiko von Beatmungsbedarf. Besonders gefährdet für Komplikationen waren ältere Patient*innen, Raucher*innen und Patient*innen, die unter COPD litten – gerade die Patient*innengruppe, die aufgrund eines erhöhten Risikos für Lungenmalignome häufig biopsiert werden muss.⁷⁹

Darüber hinaus haben Biopsien den Nachteil, die Läsion nicht stets adäquat abzubilden. Dies kann z.B. durch heterogene Wachstumsmuster der Neoplasie verursacht werden.⁸⁰ Auch lässt sich durch eine Stanzbiopsie oder Feinnadelaspiration nicht der vollständige Rundherd analysieren.⁸¹ So können sich insbesondere falsch-negative Befunde ergeben.⁸²

Eine höhere diagnostische Aussagekraft ergibt sich durch operative Verfahren. Ein besonderer Stellenwert kommt hierbei der videoassistierten thoraskopischen Chirurgie (VATS) zu. Diese minimalinvasive Chirurgie birgt die Möglichkeit, auch peripher gelegene Strukturen im Ganzen zu entfernen. Dies erlaubt, Diagnostik und Therapie in einem Arbeitsschritt zu kombinieren, während die Komplikationsrisiken im Gegensatz zu offen-chirurgischen Verfahren reduziert werden können.⁸³ Gleichzeitig besteht auch bei solch minimal-invasiven

Verfahren zusätzlich zu generellen Operationsrisiken eine höhere Komplikationsraten als bei den im vorigen Abschnitt geschilderten Biopsiemethoden.⁸⁴ Da das Hautaugenmerk dieser Studie jedoch auf radiologischer Rundherddiagnostik liegt, werden operative Verfahren an dieser Stelle nicht weiter erörtert.

Ein weiterer viel diskutierter Aspekt des derzeitigen Rundherde-Managements ist die Exposition gegenüber ionisierender Strahlung. Mit der zunehmenden Anwendung der CT ist auch die Strahlenbelastung in der Medizin angestiegen. Mettler et al. zeigten, dass sich zwischen 1950 und 2006 die Anzahl radiologischer Untersuchungen in den Vereinigten Staaten von Amerika verzehnfacht hat. Die Computertomographie macht dabei mindestens die Hälfte der gesamten effektiven Strahlendosis aus. Die pro-Kopf effektive Dosis pro Jahr durch medizinische Untersuchungen und Eingriffe hat sich zwischen 1980 und 2006 in den USA versechsfacht. Weltweit habe sich diese Dosis zwischen 1997 und 2007 nahezu verdoppelt.⁸⁵ Smith-Bindman et al. zeigten, dass im Jahre 2010 die CT zwar nur 12,0 % aller radiologischen Untersuchungen ausmachte, jedoch 67,8 % der Strahlenexposition verursachte.⁸⁶

Radiologische Strahlenexposition ist aufgrund des erhöhten Risikos von Krebserkrankungen durch ionisierende Strahlung ein kontroverses Thema. Eine der prominentesten Studien, die die Langzeiteffekte hoher Strahlenexposition untersucht haben, ist die sogenannte „Life Span Study“, die die Opfer der Atombombenangriffe auf Hiroshima und Nagasaki begleitete und einen linearen Zusammenhang zwischen Strahlendosis und Krebsrisiko beobachtete.⁸⁷ Diese Ergebnisse trugen maßgeblich zu der Entwicklung des Linear-No-Threshold-Modells (LNT) bei, das besagt, dass jegliche ionisierende Strahlung das Krebsrisiko erhöhe und dass es keine unbedenkliche Dosis gebe.⁸⁸ Dieses Modell ist bis heute der vertretene Konsens der Internationalen Strahlenschutzkommission, die für die medizinische Bildgebung eine strenge Indikationsprüfung strahlungsverursachender Untersuchungen vorsieht und empfiehlt, die Untersuchung mit der kleinstmöglichen Strahlendosis durchzuführen.⁸⁹ Diese Empfehlungen entsprechen außerdem den Vorgaben des deutschen Strahlenschutzgesetzes.⁹⁰

Im Falle der Entscheidung für die Durchführung einer CT-Untersuchung bestehen verschiedene Möglichkeiten, die Strahlendosis zu optimieren. Vor der Untersuchung spielen die korrekte Positionierung der Patient*innen, das Verwenden von Bleiabdeckungen für besonders strahlensensible Körperregionen (z.B. Mamma, Linse und Schilddrüse) und die genaue Begrenzung des Scanbereiches eine wichtige Rolle. Technische Parameter der Untersuchung wie Röhrenstrom, Rotationsdauer der Gantry, Pitch im Spiral-CT, Scandauer, Röhrenspannung und Kollimation können angepasst werden. Nach der Bilddatengewinnung greifen Bildfilter und Rekonstruktionsalgorithmen, um Artefakte und Rauschen zu reduzieren und ermöglichen scharfe Bilder trotz der vorigen Dosisreduktion.³ Die Strahlendosis eines

einzelnen CT-Scans variiert je nach untersuchter Körperregion.^{88,91} Die in den meisten Screening-Programmen für Lungenkrebs angewandte Niedrigdosis-CT des Thorax verursacht pro Scan in der Regel eine nach BMI variierende Strahlenexposition von 3-5 mSv.⁶² Dies entspricht in etwa der geschätzten natürlichen Hintergrundstrahlung in den USA pro Jahr.⁹² Andere Quellen schätzen die Dosis eines LDCT des Thorax noch niedriger ein, bei 1,4 – 2 mSv pro Scan.⁹³ Neuere Ultra-Niedrigdosis-Untersuchungen konnten die effektive Dosis noch weiter reduzieren auf ca. 0,19-0,66 mSv pro Scan.⁹⁴ Die deutsche Lungenkrebs-Früherkennungs-Verordnung gibt vor, dass für Screeningzwecke der Volumen-Computertomographie-Dosisindex nur bei adipösen Patient*innen 1,3 mGy überschreiten darf.⁹⁵

In den letzten Jahren wird das LNT teils kritisch beleuchtet und von einigen Arbeiten insbesondere im Bereich geringer Strahlendosen als zu konservativ eingeschätzt. Schultz et al. schlussfolgerten in einer 2020 veröffentlichten Metaanalyse, dass lineare Zusammenhänge von CT-untersuchungen und Krebsrisiko mehrheitlich in Fällen von deutlich höheren Strahlungsdosen gefunden wurde als durch mehrere LDCT zur Evaluation unklarer Lungenrundherde verursacht werden. Im Bereich von <100-200 mSv an kumulativer Strahlenexposition sollte kein erhöhtes Krebsrisiko bestehen.⁸⁸

Neben den periprozeduralen Risiken der Abklärung und der mit einer häufigen Nachkontrolle einhergehenden Strahlendosis ist ein weiterer wichtiger Aspekt ist die psychische Belastung, die durch eine verzögerte Rundherd-Diagnose entstehen kann. Angststörungen und Depression treten bei Rundherd-Patient*innen vermehrt auf.⁹⁶ Häufig wird ein Rundherd-Befund direkt mit einer Krebsdiagnose gleichgesetzt.⁹⁷ Dies kann zur Folge haben, dass die Patient*innen sowohl ein aggressives diagnostisches Prozedere bevorzugen.⁹⁶ Eine umfassende Aufklärung ist daher von großer Bedeutung.⁹⁸

Zusammenfassend kann festgestellt werden, dass das Management unklarer Lungenrundherde eine umfassende Thematik darstellt und insbesondere im Screening-Kontext für Lungenkarzinome zunehmend vereinheitlicht ist. Jedoch decken bestehende Richtlinien die Rundherddiagnostik im Staging-Setting und damit die besondere Fragestellung der Unterscheidung benignen und metastatischer Herde onkologischer Patient*innen nicht ab und es sind für die Dignitätsbestimmung der Rundherde weiterhin wiederholte CT-Untersuchungen, die Patient*innen ionisierender Strahlung aussetzen, und potenziell gefährliche, invasive Biopsien vonnöten. Eine Möglichkeit, mithilfe einer einzigen Bildgebungs-Untersuchung die Differenzierung zwischen benignen und malignen bzw. metastatisch sicher durchzuführen, gibt es zurzeit noch nicht. Fortschritte im Feld der künstlichen Intelligenz ermöglichen jedoch neue Verfahren zur Analyse und Verarbeitung von Bilddaten, um die

Diagnostik zu vereinfachen und zu beschleunigen. Dazu gehört das Feld der sogenannten Radiomics. Diese werden im Folgenden erläutert.

2.3. Radiomics

2.3.1. Hintergrund

Der Ansatz, die radiologische Diagnostik mithilfe moderner Technologien zu vereinfachen, zu optimieren oder gar vollständig zu automatisieren, stellt keine neuartige Idee dar. Nach ersten erfolglosen Versuchen der Quantifizierung radiologischer Bilddaten in den 1960er-Jahren, waren maschinelle Lernprozesse und Mustererkennungsalgorithmen in den 1980er-Jahren erstmalig so weit fortgeschritten, dass computerassistierte Diagnostik angewandt werden konnte, zunächst in der Diagnostik von Mamma- und Lungenkarzinomen sowie kardiovaskulärer Erkrankungen.⁹⁹

Mit der fortschreitenden Digitalisierung entstehen insbesondere in den letzten Jahrzehnten stetig wachsende Mengen an Daten.¹⁰⁰ Dieses Phänomen wird auch als *big data* bezeichnet und hat auch in der Medizin bzw. der medizinischen Forschung Einzug gehalten.¹⁰¹ In Zusammenhang mit Fortschritten im Bereich der künstlichen Intelligenz und des maschinellen Lernens haben sich zahlreiche methodische Ansätze zur Analyse dieser Daten etabliert. Das Suffix -omics wird verwendet, um solche Prozesse zu beschreiben.¹⁰⁰ In der medizinischen Bildgebung hat sich so der Begriff Radiomics etabliert. Dieser wurde 2012 geprägt¹⁰² und beschreibt die quantitative Analyse radiologischen Bildmaterials verschiedener Untersuchungsmodalitäten.¹⁰⁰ Radiomics-Analysen werden mehrheitlich an CT-, MRT- oder PET-Daten durchgeführt, sind aber auch auf andere Modalitäten übertragbar.¹⁰³

Grundlage von Radiomics sind die sogenannten Features. Diese quantitativen Bildcharakteristika werden über verschiedene Algorithmen den Bilddaten extrahiert.¹⁰³ Features lassen sich in verschiedene Kategorien einteilen: Shape-Features beschreiben geometrische Eigenschaften wie Volumen und Oberfläche. First-Order-Features, auch Histogramm-basierte Features genannt, beschreiben die Verteilung der Intensitäten innerhalb der untersuchten Struktur. Dazu gehören u.a. Median, Standardabweichung, Kurtosis und Entropie (statistischer Informationsgehalt).¹⁰⁴ Second-Order-Features, auch Texture-Features genannt, stellen die statistische Beziehung der Intensitäten zweier benachbarter Voxel (dreidimensionale Pixel) oder Voxelgruppen dar und dienen daher der Quantifizierung der Heterogenität innerhalb der untersuchten Struktur.¹⁰⁵ Dazu gehören u.a. verschiedene Graustufen-Matrizen.¹⁰⁶ Letztlich gibt es sogenannte Higher-Order-Features. Für diese müssen die Bilddateien zunächst durch mathematische Transformationen oder Filter prozessiert werden, um Rauschen zu reduzieren und so bestimmte Muster sicht- und quantifizierbar zu machen. Darunter fallen z.B. Gauß-Filter sowie Wavelet- oder Laplace-Transformationen.¹⁰⁴

Mithilfe dieser Features als quantitative Charakteristika können Modelle entwickelt werden, die verschiedene medizinische Fragestellungen erörtern.

Vorteil von Radiomics ist, dass die Erfassung von Bildeigenschaften möglich ist, die dem menschlichen Auge nicht zugänglich sind.¹⁰⁷ Es können so Rückschlüsse auf klinisch relevante Parameter wie zelluläre¹⁰⁸ und genetische Heterogenität¹⁰⁹ oder Tumoraggressivität¹¹⁰ gezogen, neue Biomarker entdeckt¹¹¹ und prädiktive Modelle sowohl für Diagnostik¹⁰⁰ als auch für Prognoseeinschätzungen entwickelt werden.¹¹² Darüber hinaus können auch sichtbare Eigenschaften quantifiziert werden, um deren Beurteilung zu standardisieren und Variabilität durch verschiedene Radiolog*innen zu reduzieren.¹⁰⁰

Auf diese Art und Weise sind tiefergehende Analysen von z.B. unklaren Raumforderungen möglich, ohne die Risiken und Kosten einer Biopsie oder die Strahlenbelastung in zahlreichen Folgeuntersuchungen eingehen zu müssen.¹¹³ Außerdem lassen sich Raumforderungen im Ganzen analysieren und nicht nur teilweise wie z.B. durch Stanzbiopsien.¹¹¹

Letztendlich ergeben sich mithilfe von Radiomics durch den Sprung von qualitativer zu quantitativer Bildgebung¹¹³ neue Möglichkeiten personalisierter Diagnostik und Therapie.¹¹⁴

2.3.2. Künstliche Intelligenz, maschinelles Lernen und Deep Learning

Im Rahmen von Radiomics-Prozessen müssen große Datenmengen analysiert und diese auf die entsprechenden Fragestellungen angewandt werden. Hierfür kommen verschiedene Methoden von maschinellem Lernen (ML) und künstlicher Intelligenz (KI) zur Anwendung. Aus diesem Grund folgt an dieser Stelle ein Überblick über diese Thematik.

Künstliche Intelligenz gewinnt in den letzten Jahren stetig an Bedeutung. Der Begriff wurde erstmalig 1956 geprägt und beschreibt die Möglichkeit, intellektuelle Aufgaben, die vorher Menschen vorbehalten waren, zu automatisieren.¹¹⁵ Daraus resultieren vielfältige Entwicklungsperspektiven sowohl in wissenschaftlichen als auch in nicht-wissenschaftlichen Anwendungsbereichen.¹¹⁶ Die sogenannte symbolische KI basiert auf der Erlernung klar definierter Regeln für sämtliche mögliche Ausgänge einer Situation oder Fragestellung und trifft Entscheidungen auf Grundlage dieser Regeln. Diese Form der KI wird meist zur Lösung logischer, mathematischer Probleme angewandt.¹¹⁶ Für Aufgaben, die ein höheres Maß an Abstraktion und Mustererkennung erfordern, ist diese Form der KI jedoch ungeeignet. Stattdessen werden hierfür Methoden des maschinellen Lernens eingesetzt. Dabei kommen Analyseverfahren sowie spezifische Algorithmen zum Einsatz, um statistische Muster in Datensätzen zu identifizieren und daraus Vorhersagen für bislang unbekannte Daten abzuleiten.¹¹⁷ Für eine optimale Modellbildung ist ein ausreichend großer Datensatz erforderlich, der das Training der entsprechenden Algorithmen ermöglicht.¹¹⁸

Maschinelle Lernprozesse können in vier verschiedene Klassen unterteilt werden: überwachtes Lernen, unüberwachtes Lernen, halb-überwachtes Lernen und bestärkendes Lernen.¹¹⁶

Beim überwachten Lernen wird der Datensatz typischerweise in Trainings- und Testdaten unterteilt. Die Trainingsdaten sind annotiert, sodass der Algorithmus lernt, die Eingaben den jeweiligen Zielvariablen korrekt zuzuordnen.¹¹⁹ Diese Methodik wird insbesondere für Regressionsaufgaben, wo numerische Vorhersagen getroffen werden, und für Klassifikationsaufgaben verwendet.¹¹⁶ Aufgrund der erforderlichen Annotierung der Daten ist dieser Prozess jedoch mit erheblichem Zeit- und Kostenaufwand verbunden.¹²⁰

Beim unüberwachten Lernen fehlt die Annotierung der Daten, sodass der Algorithmus nicht nur eigene Muster erfasst, sondern daraufhin auch eigene Kategorien als Ausgabedaten generiert. Diese Methode eignet sich insbesondere für die Detektion von Anomalien, für Assoziationsaufgaben sowie für die Clusterbildung anhand spezifischer Merkmalskombinationen.¹¹⁶

Eine Mischform der oben erläuterten Lernprozesse stellt das halb-überwachte Lernen dar, bei dem lediglich ein Teil der Daten annotiert ist. Das an den annotierten Daten trainierte Modell wird anschließend auf die nicht annotierten Daten angewandt.¹²⁰

Das sogenannte bestärkende Lernen unterscheidet sich von den vorigen Lernprozessen, da es auf dem Versuch-und-Irrtum-Prinzip (eng: trial and error) basiert. Jede Aktion des Algorithmus wird positiv oder negativ bewertet, bis über eine erlernte Abfolge von Entscheidungen das gewünschte Ergebnis erreicht ist. Die einzelnen Aktionen sind dabei nicht strikt richtig oder falsch, vielmehr wird das Ziel durch die Optimierung einer Abfolge von Entscheidungen angestrebt.¹²¹ In der Medizin wird diese Technik bisher nur selten angewandt.¹¹⁶

Es gibt eine Vielzahl unterschiedlicher statistischer Analyseverfahren bzw. ML-Algorithmen, die Anwendung finden. Regressionsanalysen, die überwiegend im überwachten Lernen eingesetzt werden, erlauben die Identifikation von Zusammenhängen zwischen Variablen und ermöglichen auf dieser Basis Vorhersagen und dadurch letztlich Entscheidungen. Für lineare Zusammenhänge zwischen kontinuierlichen Variablen wird lineare Regression verwendet. Je nach Anzahl dieser Variablen wird univariate oder multivariate lineare Regression angewandt, beispielweise um die verschiedenen Einflussfaktoren auf den Preis einer Immobilie (Standort, Größe und Ausstattung etc.) darzustellen.¹²²

Logistische Regression analysiert Zusammenhänge kategorischer Variablen (z.B. die Unterscheidung der Dignität eines Lungenrundherds) und gibt numerische Wahrscheinlichkeiten spezifischer Ergebnisse an. Verschiedene Verfahren kommen zur Anwendung. Entscheidungsbäume teilen die Eingabedaten wiederholt nach verschiedenen Parametern auf, bis die finale Vorhersage oder Entscheidung getroffen wird. Im Rahmen von

K-Nearest-Neighbour-Algorithmen werden die einzelnen Datenpunkte in verschiedene Klassen unterteilt. Die Vorhersagen für neue Datenpunkte erfolgen abhängig davon, in welcher Klasse die ähnlichsten bzw. nächsten Datenpunkte im zweidimensionalen Raum liegen. Weitere verbreitete Algorithmen sind Support Vector Machine, Gradientenabstieg, Bayesian Learning Naive Bayes, K-Means-Clustering und Backpropagation.¹²²

Im Rahmen von komplexeren Methoden maschinellen Lernens können so genannte künstliche neuronale Netze (KNN) Anwendung finden. Hierbei sollen biologische neuronale Netze imitiert werden. Diese bestehen aus einzelnen Neuronen, die regelmäßig von anderen Neuronen Eingaben in Form von Aktionspotentialen erhalten und basierend auf den Charakteristika dieser Eingaben binär reagieren: Entweder wird ein Aktionspotential generiert oder nicht. Diese Reduktion der Komplexität einer Vielzahl eingehender Daten durch Kategorisierungen der Input-Muster ist das Ziel künstlicher neuronaler Netze.¹²³ Ein künstliches Neuron, das Eingaben summiert, gewichtet, und unter Anwendung einer Aktivierungsfunktion binär reagiert, ist der Grundbaustein von KNN¹²³ und wird als Perceptron bezeichnet.¹¹⁶ Diese Grundbausteine bilden untereinander sogenannten Schichten, die jeweils lineare Probleme lösen. Sind die Ziel-Klassen jedoch nicht linear zu trennen, werden mehrschichtige neuronale Netzwerke genutzt. Zwischen Eingabe- und Ausgabeschicht werden hier eine oder mehrere verdeckte Schichten (eng: hidden layers) geschaltet. Dadurch wird die Modellierung komplexer nichtlinearer Zusammenhänge ermöglicht. Mit zunehmender Anzahl an Schichten steigt die Komplexität der Analysen.¹²³ Modelle, die sehr viele verborgene Schichten enthalten, werden als Deep-Learning-Modelle bezeichnet. Diese eignen sich besonders zur Analyse komplexer Eingabedaten wie z.B. radiologischem Bildmaterial.¹¹⁷

Maschinelle Lernprozesse sind allesamt abhängig von sogenannten Hyperparametern. Dabei handelt es sich um Variablen, die das Lernverhalten und dadurch die Ergebnisse eines entstehenden Modells beeinflussen, z.B. die Anzahl der Schichten eines komplexen neuronalen Netzes oder der Anzahl der Verzweigungen in Entscheidungsbaum-basierten ML-Algorithmen. Die Auswahl der optimalen Hyperparameter erfolgt heutzutage zunehmend automatisiert über spezielle Software.¹²⁴

2.3.3. Allgemeiner Arbeitsablauf

Der Arbeitsablauf eines Radiomics-Prozesses beginnt mit der Gewinnung des radiologischen Bildmaterials. Im nächsten Schritt erfolgt die Segmentierung, bei der die Region of Interest (ROI) im zweidimensionalen bzw. das Volume of Interest im dreidimensionalen Bereich präzise konturiert wird.¹⁰³ Dieser Segmentierungsprozess kann rein manuell, aber auch semi- oder vollautomatisiert mithilfe verschiedener Algorithmen erfolgen.¹⁰⁰ Automatisierte Prozesse

sind zeiteffizienter und tragen zugleich zur Reduktion der potenziellen Variabilität bei, die bei manuellen oder semi-automatisierten Segmentationsprozessen durch unterschiedliche Untersucher*innen entstehen kann. Die Generalisierbarkeit der Algorithmen zur automatischen Segmentierung ist derzeit jedoch noch begrenzt.¹⁰⁰

Vor der Extraktion der Features erfolgt eine Prozessierung der segmentierten Bilddaten, um einen möglichst homogenen Datensatz zu erhalten. Zu den gängigen Techniken gehören u.a. die Vereinheitlichung von Voxelgröße und -abstand aller Bilddateien, der Ausschluss von Voxel, die außerhalb individuell angepasster Grenzwerte liegen, um z.B. Luft oder Knochengewebe auszuschließen und im Falle von MRT-Bildern die Quantifikation der Intensitäten.¹⁰⁰ Viele dieser Prozesse können mithilfe verschiedener Software, z.B. PyRadiomics¹²⁵, automatisiert werden. Daraufhin werden die Features mithilfe verschiedener Algorithmen der entsprechenden Radiomics-Software berechnet und extrahiert.¹⁰³

In der Regel wird nur ein Teil der extrahierten Features zur Erörterung einer wissenschaftlichen Fragestellung verwendet. Im Falle von Klassifikationsmodellen birgt eine hohe Anzahl an Features, insbesondere in Kombination mit wenigen Ziel-Klassen, das Risiko für ein Overfitting (Überanpassung) der aus den Features entwickelten Modelle auf den Trainingsdatensatz. Dadurch sinkt die Generalisierbarkeit des Modells.¹⁰⁰ Daher wird für die Modellentwicklung die Feature-Anzahl reduziert. Dieser Prozess wird als Selektion bezeichnet. Hierbei werden Merkmale ausgeschlossen, die entweder nicht robust sind, das heißt unter Variation der Untersuchungsparameter instabil sind¹²⁶, nicht reproduzierbar, also bei Durchführung durch andere Personen oder in anderen Zentren trotz identischer Methodik und Datensätze keine konsistenten Ergebnisse liefern¹²⁶, oder redundant, also stark korreliert, sind.¹⁰⁰ Zur Identifizierung der relevantesten Features für die jeweilige Fragestellung kommen verschiedene Metriken und Algorithmen des maschinellen Lernens zum Einsatz (siehe 2.3.2.).

Anschließend können die ausgewählten Features genutzt werden, um Modelle zur Lösung der Fragestellung zu entwickeln. Auch hier kommen die o.g. Methoden des maschinellen Lernens oder des Deep Learnings zur Anwendung.¹⁰³ Die Auswahl der Methoden variiert je nach Studie. Nach Erstellung eines Modells ist dessen Validierung erforderlich, die entweder durch externe Datensätze, welche unabhängig von den zur Modellerstellung verwendeten Daten sind, oder durch das interne Verfahren der Kreuzvalidierung erfolgen kann. Letztere beinhalten die Aufteilung des Datensatzes einer Studie in Subgruppen, üblicherweise in sogenannte Trainingsdatensätze und Testdatensätze. Auf diese Weise kann das Modell unter Verwendung des Trainingsdatensatzes entwickelt und anschließend an den Testdatensätzen überprüft werden, ohne dass hierfür die Gesamtheit der erhobenen Daten herangezogen werden muss. Dadurch wird das Risiko des bereits erwähnten Overfittings reduziert. Die Validierung erfolgt also anhand von Daten, die während des Trainingsprozesses unberührt geblieben sind,

sodass die Verwendung eines zusätzlichen externen Datensatzes nicht zwingend erforderlich ist.¹²⁷ Eine erweiterte Version klassischer Kreuzvalidierung stellt die sogenannte verschachtelte Kreuzvalidierung (eng: nested cross validation) dar. Hierbei werden einzelne der zuvor gebildeten Datensätze nochmals unterteilt, sodass eine innere und eine äußere Schleife entstehen (siehe Abbildung 1). Innerhalb der Datensätze der inneren Schleife erfolgt eine unabhängige Optimierung der Hyperparameter des Modells. Erst nach Abschluss dieses Schrittes wird das optimierte Modell anhand der in der äußeren Schleife enthaltenen Datensätze getestet.¹²⁷

2.3.4. Limitationen

Trotz des großen Potenzials von Radiomics weist die Technik derzeit noch verschiedene Limitationen auf.

Um die Studienbedingungen z.B. für externe Validierungsprozesse nachzubilden und evaluieren zu können, sind detaillierte Angaben zu Bildgebungsprotokollen, Bildprozessierung sowie den eingesetzten Berechnungsalgorithmen der Features erforderlich. Diese sind jedoch aufgrund insuffizienter Dokumentation in Publikationen oft nicht gegeben. Dies führt dazu, dass viele der Studien nicht extern validiert sind und daher deren klinische Anwendung noch begrenzt ist.¹⁰⁰ Darüber hinaus sind die meisten Radiomics-Studien retrospektiver Natur, was u.a. ein erhöhtes Risiko für Selektionsbias und grundsätzlich einen geringeren Evidenzgrad mit sich bringt.¹⁰⁰ Der Radiomics Quality Score (RQS) wurde entwickelt, um die Dokumentation des Studiendesigns zu vereinheitlichen und die Qualität der Studien zu evaluieren. Er enthält 16 Punkte, die verschiedene Aspekte, z.B. Segmentierung und Validierung erfassen.¹²⁸

Einen zentralen limitierenden Faktor stellt die Robustheit der extrahierten Features dar. Verschiedene technische Rahmenbedingungen der Bildakquisition und -verarbeitung beeinflussen die Quantifizierung maßgeblich. So konnte gezeigt werden, dass Features zwischen unterschiedlichen CT-Scannern erheblich variieren. Lennartz et al. verglichen den Effekt von drei verschiedenen Dual-Energy-CT-Geräten auf die Features in einer Patient*innenkohorte und an einem Phantom-Modell. Keines der am Phantom gewonnenen Features zeigte einen Konkordanz-Korrelationskoeffizienten (CCC) von $\geq 0,90$. In der Patient*innenkohorte konnten maximal 16,15 % diesen Wert erreichen.¹²⁹

Auch die Schichtdicke spielt eine wichtige Rolle. Park et al. verglichen die Robustheit der Features von drei verschiedenen Schichtdicken: 1 mm, 3 mm, 5 mm. Die CCC zwischen den verschiedenen Schichtdicken lagen nur bei 0,27 – 0,65. Nach Konversion der zwei dickeren Schichten in 1-mm-Schichten mithilfe eines Deep-Learning-Algorithmus stiegen die CCC auf 0,45-0,72 an.¹³⁰

Rohdaten einer Schnittbildgebungsmodalität können mit einer Vielzahl an Algorithmen, Filtern und Kernels rekonstruiert werden, um auswertbare Schnittbilder zu generieren. Diese variieren zwischen verschiedenen Herstellern und werden z.T. auch in Kliniken individuell angepasst.¹³¹ Dies führt zu einer Variabilität der Features.¹³² Weitere Einflussfaktoren sind u. a. Röhrenspannung, Röhrenstrom, Pitch-Faktor, Field-of-View¹⁰⁰ und Voxel-Größe.¹³³

Um der Problematik der mangelnden Robustheit, Reproduzierbarkeit und Variabilität der Features entgegenzuwirken, wurden in den letzten Jahren Initiativen zu deren Standardisierung ergriffen. Die Image Biomarker Standardization Initiative (IBSI) aus dem Jahre 2020 konnte mittels klar definierter Nomenklatur und Bildprozessierungsschritten 169 Features standardisieren.¹¹⁴ Darüber hinaus wurden verschiedene Kompensationsmethoden entwickelt, um den Einfluss der oben genannten Faktoren auf die Features zu reduzieren.¹³⁴⁻

136

Eine weitere Limitation vieler Radiomics-Studien sind Intra-und Interreader-Variabilität. Die Segmentierung der ROI stellt den wichtigsten und anspruchsvollsten Abschnitt des Arbeitsablaufs einer Radiomics-Studie dar, da sie maßgeblich die Features, die extrahiert werden, bestimmt.¹⁰⁵ Derzeit erfolgen die Segmentierungen noch mehrheitlich manuell oder semi-automatisiert. Dies birgt die Gefahr uneinheitlicher Segmentierungen und folglich nicht robuster oder nicht reproduzierbarer Features, sowohl in Studien mit mehreren Segmentierenden/Readern als auch in Studien mit nur einer segmentierenden Person.¹³⁷

Maschinelle Lernprozesse, die die Grundlage von Radiomics-Modellen darstellen, sind gänzlich abhängig von den Eingabedaten. Barragán-Montero et al. beschreiben zwei Kategorien der schlechten Eingabedaten: insuffiziente Daten (quantitativ) und unpassende Daten (qualitativ).¹³⁸ Die Kategorie der insuffizienten Daten beinhaltet Datensätze, die zu klein sind, um Patient*innenvariablen umfassend darzustellen. Unpassende Daten dagegen beinhalten verschiedene Faktoren: Dazu gehören unbalancierte Datensätze, wo einige Klassen innerhalb des Datensatz häufiger repräsentiert sind als andere, was zu falschen Klassifikationen führen kann, weil das Modell in Richtung der überrepräsentierten Klasse verzerrt wird. Der Effekt eines solchen Klassenungleichgewichts (eng: class imbalance) kann z.B. durch Vergrößerung oder Verkleinerung der entsprechenden Klassen des Trainings-Datensatzes (Over/Undersampling) minimiert werden.¹³⁹ Weitere qualitative Probleme im Kontext von Radiomics können Artefakte und Rauschen der Bilddaten, minderqualitative Annotationen und bereits erwähnte Interreader-Variabilität sein.¹³⁸

Ein weiterer häufig diskutierter Aspekt, der im Kontext der Verwendung maschineller Lernprozesse zur Sprache kommt, ist das sogenannte Black-Box-Phänomen. Da diese Lernprozesse nach entsprechender Programmierung automatisiert ablaufen, lassen sich die

einzelnen Schritte, die unternommen wurden, um auf die Ergebnisse bzw. Ausgabedaten zu kommen, nicht nachvollziehen, was die Interpretation dieser Ergebnisse erschwert.¹³⁸

Trotz ihres großen Potenzials finden Radiomics weiterhin kaum klinische Anwendung. Diese translationale Lücke ist maßgeblich durch die in diesem Abschnitt geschilderten Limitationen wie Feature-Reproduzierbarkeit/Robustheit, unzureichende externe Validierung und mangelnde Interpretierbarkeit der Ergebnisse bedingt. Außerdem müssen die entsprechenden Softwares komplexe regulatorische Prozesse durchlaufen, bevor sie für die klinische Anwendung zugelassen werden.¹⁴⁰

2.4. Anwendung von Radiomics

2.4.1. Anwendungen in der Onkologie

Radiomics finden in verschiedensten medizinischen Fachbereichen Anwendung. Dazu gehören unter anderem die Früherkennung der diabetischen Retinopathie¹⁴¹, die Subtypisierung der frontotemporalen Demenz¹⁴², die Evaluation des Blutungsrisikos ösophagealer Varizen¹⁴³ oder die Unterscheidung chronisch entzündlicher Darmerkrankungen.¹⁴⁴

Mehrheitlich haben Radiomics bisher jedoch in der Onkologie Anwendung gefunden, sowohl im diagnostischen als auch im therapeutischen Bereich.^{145,146} Neben der Klassifikation von benignen und malignen Strukturen¹⁴⁷, welche am Beispiel von Lungenrundherden unter Punkt 2.4.2. ausführlich besprochen wird, ergibt sich eine Vielzahl anderer Anwendungsmethoden.

Klassifikationsstudien zu onkologischen Fragestellungen verwenden neben weit verbreiteten statistischen Metriken wie Sensitivität und Spezifität weitere statistische Kennziffern zur Evaluation der Klassifikationsleistung von Modellen. Dazu gehört die Balanced Accuracy, die den Durchschnitt der korrekt klassifizierten Fälle der verschiedenen vertretenen Klassen darstellt. Dies verhindert Verzerrung der Ergebnisse bei Datensätzen mit Klassenungleichgewichten.¹⁴⁸ Der F1-Score stellt das arithmetische Mittel aus Sensitivität/Recall und positivem prädiktivem Wert/Präzision dar.¹⁴⁸ Eine weitere Metrik stellt die Area Under the Receiver Operator Characteristic (ROC) Curve (AUROC, kurz AUC) dar. Diese erlaubt eine graphische Darstellung der Klassifikationsleistung im zweidimensionalen Raum. Hierbei wird für verschiedene Schwellenwerte das Ergebnis der Klassifikation anhand von Sensitivität und (1-Spezifität) erfasst. Anhand der Schwellenwerte lässt sich so eine Kurve, die sogenannte ROC, konstruieren. Die Fläche unter dieser Kurve, die AUC, gilt als Maß der Klassifikationsfähigkeit.¹⁴⁹ Bei sämtlichen vorgestellten Kennziffern stellt ein Wert von 1,0 eine optimale Klassifikation dar.

Darüber hinaus kommen häufig statistische Nomogramme zur Anwendung. Diese sind in der Lage, die verschiedenen Variablen eines Modells (z.B. klinische und Radiomics-basierte Parameter eines Klassifikationsmodells) samt deren individueller Wahrscheinlichkeiten und Einflüsse auf das Gesamtergebnis des Modells grafisch darzustellen.¹⁵⁰

Radiomics-Features können zur nicht-invasiven Typisierung von Krebserkrankungen herangezogen werden. Dazu gehört beispielsweise die Unterscheidung pulmonaler Adenokarzinome und Plattenepithelkarzinome¹⁵¹ oder die Unterscheidung verschiedener Neoplasien des Magens.^{152,153} Des Weiteren konnten in Radiomics-Studien auch therapie- und prognoserelevante Mutationen identifiziert werden, z.B. EGFR-Mutationen in Hirnmetastasen¹⁵⁴ oder KRAS/NRAS/BRAF-Mutationen in Kolorektalkarzinomen.¹⁵⁵ Außerdem ließ sich die Invasivität von Adenokarzinomen der Lunge abschätzen.¹⁵⁶

Radiomics können außerdem zum Tumor-Staging eingesetzt werden. Die extrahierten Features können Hinweise auf das Vorliegen von Lymphknotenmetastasen liefern und stellen potenziell eine nicht-invasive Alternative zur Wächterlymphknotenbiopsie beim Mammakarzinom dar.¹⁵⁷ Qi et al. konnten außerdem zeigen, dass mittels Radiomics bereits Hinweise auf zukünftige Metastasierung entdeckt werden können, bevor diese auf CT-Bildern sichtbar werden. Es konnten statistische Unterschiede der Features zwischen prä-metastatischen und nicht-metastatischen Lungenregionen gezeigt werden, die ggf. der bereits beginnenden Tumorerkrankung entsprechen. Bezüglich der Unterscheidung metastatischer und nicht-metastatischer Regionen konnten eine Sensitivität und Spezifität von je 92,9 % gezeigt werden.¹⁵⁸

Einen weiteren Schwerpunkt bilden Prognosemodelle. So konnte gezeigt werden, dass die Kombination klinischer Prognosefaktoren (z.B. Tumorstadium und Tumolvolumen) mit einem Radiomics-Score eine höhere Genauigkeit bei der Überlebensvorhersage von Patient*innen mit High-grade-Osteosarkomen erzielt als die klinischen Faktoren allein.¹¹² Coroller et al. konnten zudem mittels Features, die aus Adenokarzinomen der Lungen gewonnen wurden, das Fernmetastasierungsrisiko einschätzen.¹⁵⁹

Auch eine Beurteilung des Therapieerfolgs mittels Radiomics ist möglich. Quantitative Analysen von T2-gewichteten MRT-Untersuchungen zeigten eine präzisere Analyse des lokalen Therapieerfolgs nach Radiochemotherapie bei Rektumkarzinom als subjektive Beurteilung durch erfahrene Radiolog*innen.¹⁶⁰ In einer Studie von Mattonen et al. zeigte ein Radiomics-basiertes Modell ebenfalls bessere Ergebnisse bei der Früherkennung von Lokalrezidiven von Lungenkarzinomen nach stereotaktischer Bestrahlungstherapie als die Beurteilung durch erfahrene Radiolog*innen.¹⁶¹

2.4.2. Radiomics zur Klassifikation unklarer Lungenrundherde

Die Klassifikation unklarer Rundherde ist der Fokus zahlreicher Radiomics-Studien. Die Unterscheidung benigner Rundherde von primären Lungenkarzinomen ist bereits breit erforscht.

Hawkins et al. konnten auf Grundlage der Daten des NLST ein Klassifikationsmodell entwickeln, das 23 verschiedene Features verwendete und eine AUC von bis zu 0,83 für die Unterscheidung benigner und (prä-)maligner Läsionen erreichte.¹⁶² Warkentin et al. nutzen neben Radiomics-Features, die aus segmentierten Rundherden mehrerer Screeningstudien gewonnen wurden, klinisch-epidemiologische Variablen (z.B. Alter, Raucherstatus, Familienanamnese) und kombinierten diese zu einem Klassifikationsmodell, das bessere Ergebnisse zeigte als das unter 2.2.1.2. erwähnte PanCan-Modell (AUC bis 0,93).¹⁶³ Ein ähnliches Studiendesign nutzten Liu et al. Ihr Modell kombinierte Radiomics-Features mit der Variable des Patient*innenalters zur Differenzierung von Adenokarzinomen und benignen Rundherden. Die AUC lag hier bei bis zu 0,987.¹⁶⁴ Zhao et al. entwickelten ein Klassifikations-Nomogramm zur Unterscheidung von Adenokarzinomen und Kryptokokkoseherden, das 24 Features zusammen mit den Faktoren Durchmesser, Lobulierung und pleuralen Einziehungen zu einem Nomogramm vereinte, zeigte eine AUC von 0,8881. Das Radiomics-Modell allein zeigte eine AUC von 0,8306, höher als die der anderen Komponenten des Nomogramms (0,7901).¹⁶⁵

Neben klinischen Variablen kann auch das umliegende Lungenparenchym genutzt werden, um die Klassifikation unklarer Rundherde zu ergänzen. Beig et al. entwickelten ein Modell, das intranoduläre und perinoduläre Features kombinierte und in der Unterscheidung von histopathologisch bestätigten Granulomen und Adenokarzinomen eine AUC von 0,80 erreichte.¹⁶⁶ Dilger et al.¹⁶⁷ und Yi et al.¹⁶⁸ beschrieben ebenfalls eine Verbesserung der Klassifikationsergebnisse unter Verwendung des umliegenden Lungenparenchyms.

Deutlich weniger breit untersucht ist die Unterscheidung benigner Rundherde von Metastasen mittels Radiomics.¹⁶⁹ Özgül et al. klassifizierten Metastasen von Harnblasenkarzinomen, mittels eines Radiomics-Modells auf Basis von Histogramm-basierten Features. Hier konnte eine AUC von 0,702 erreicht werden.¹⁷⁰ Cho et al. untersuchten Osteosarkom-Metastasen in pädiatrischen Patient*innen. Das Radiomics-basierte Klassifikationsmodell zeigte eine AUC von 0,865.¹⁷¹ Hu et al. verglichen ein klinisches Modell (unter Einbezug von N-Stadium, Chronizität und Herdgröße) mit einem Radiomics-Modell sowie einer kombinierten Variante in Form eines Nomogramms zur Diagnostik von Lungenmetastasen von Kolorektalkarzinomen. Während das klinische Modell eine AUC von 0,852 erreichte, erzielte das Radiomics-Modell 0,881. Das kombinierte Nomogramm wies mit einer AUC von 0,922 die höchste Genauigkeit auf.¹⁷²

Die oben genannten Studien zur Klassifikation von Lungenmetastasen zeigen zwar überzeugende Ergebnisse, sind jedoch meist auf die Metastasen einer einzelnen onkologischen Grunderkrankung spezialisiert und haben entsprechend kleine Studienkohorten.^{170,171} Dies vermindert jedoch die Generalisierbarkeit der entwickelten Modelle auf ein breites Patient*innenkollektiv. Lennartz et al. nutzten in ihrer Studie zur Differenzierung benigner und metastatischer Rundherde sowohl konventionelle CT-Aufnahmen als auch Iodkarten, die mittels Dual-Energy-CT (DECT) gewonnen wurden. Hier wurde bewusst eine heterogene Kohorte onkologischer Patient*innen mit zahlreichen Grunderkrankungen und insgesamt 571 Rundherden eingeschlossen. Die Kombination der stärksten Features beider Aufnahmen ergab eine Genauigkeit von 0,87. Rein konventionelle CT-Features dagegen zeigte eine nicht statistisch signifikant verminderte Genauigkeit (0,84).⁶⁴

Ein weiterer Faktor, der die breite Anwendbarkeit der bisherigen Klassifikationsmodelle einschränkt besteht darin, dass viele der o.g. Studien nur eine oder wenige CT-Scanner mit einheitlich Bilddatengewinnungsparametern und Rekonstruktionsalgorithmen verwenden.^{64,81,173} Während verschiedene Scanner mit individuellen Untersuchungsparametern negative Effekte auf die gewonnenen Features haben können (siehe 2.3.4.)¹²⁹, spiegelt eine Homogenität der Untersuchungsbedingungen jedoch den klinischen Alltag nur unzureichend wieder und die Generalisierbarkeit des Modells wird eingeschränkt.⁶⁴

2.5. Fragestellungen und Ziel der Arbeit

Die große Mehrheit der bisherigen Studien, die sich mit der Klassifikation unklarer Lungenrundherde mittels Radiomics beschäftigen konzentriert auf die Unterscheidung primärer Lungenmalignome von benignen Rundherden. Außerdem erfolgen diese Studien meist in kleineren Kohorten und beziehen nur Bilddaten einzelner oder weniger Scanner mit ein.¹⁶⁸ Studien zur expliziten Erfassung von pulmonalen Metastasen sind dagegen selten beschränken sich darüber hinaus meist auf Metastasen einer einzelnen onkologischen Grunderkrankung.¹⁷¹ Auf diese Art und Weise wird jedoch die Generalisierbarkeit der Ergebnisse reduziert. Im klinisch-radiologischen Alltag, insbesondere im Kontext mehrjähriger Folgeuntersuchungen, ist es eher Regel als Ausnahme, dass ein breites Spektrum onkologischer Patient*innen Untersuchungen an verschiedenen Geräten mit unterschiedlichen Untersuchungsparametern erhalten.

Ziel dieser Studie war es daher, die Unterscheidung benigner Lungenrundherde von Lungenmetastasen mittels eines Radiomics-Modells durchzuführen, das an einer großen Kohorte mit ausgeprägter Heterogenität bzgl. onkologischer Erkrankungen und Untersuchungsbedingungen erstellt wurde, um eine größtmögliche Anwendbarkeit für die

klinische Praxis zu erreichen und gleichzeitig den Einfluss der besagten Heterogenität einschätzen zu können.

3. Material und Methoden

3.1. Kohorte

3.1.1. Einschlusskriterien der Proband*innen

Potenziell geeignete Patient*innen für die Studie wurden mithilfe einer Durchsuchung der radiologischen Datenbank der Uniklinik Köln vom 01.05.2010 bis zum 31.07.2020 identifiziert. Dabei wurden folgende Suchparameter verwendet:

- ≥18 Jahre alt
- in onkologischer Behandlung
- kontrastmittelgestützte Computertomographie des Thorax in venöser Phase
- Befunde enthalten folgende Stichworte:
 - Metastasen: „Metastasen“ und entsprechende Synonyme (z.B. „Filiae“)
 - Benigne Rundherde: „benigne[r] Rundherd[e]“, „Knoten“/„Knötchen“ „Granulom[e]“, weitere benigne Differenzialdiagnosen

Die Resultate der Datenbanksuche wurden zusammengefasst und im nächsten Schritt wurden die Untersuchungen einzeln darauf überprüft, ob passende Lungenrundherde vorliegen. Ein Nichterfüllen der unten genannten Einschlusskriterien führte zum Ausschluss.

3.1.2. Einschlusskriterien der Lungenrundherde

Zum Einschluss in die Studie das Vorliegen einer verlässlichen Ground Truth, also eines Referenzstandards, hinsichtlich der Dignität erforderlich. Hierfür galten folgende Kriterien:

Benigne Rundherde:

1. Histopathologische Bestätigung der Benignität und/oder
2. Stabiler Durchmesser in mehreren Computertomographie-Untersuchungen über einen Zeitraum von mindestens 6 Monaten ohne antineoplastische Systemtherapie

Metastasen:

1. Histopathologische Bestätigung der Metastase und/oder
2. Malignomtypische Radionuklid Aufnahme in der PET-CT-Bildgebung und/oder
3. Größenzunahme von mindestens 20 % im Vergleich zu früheren oder nachfolgenden CT-Untersuchungen
4. Größenreduktion oder vollständiges Verschwinden des Rundherds in nachfolgenden CT-Untersuchungen unter antineoplastischer Therapie

Ungeachtet der Dignität wurde zudem für die eingeschlossenen Läsionen ein Minestdurchmesser von 5 mm festgelegt. Deren Segmentierung ist insbesondere im Bereich

<5 mm anspruchsvoll, da einzelne Voxel im Randbereich neben relevantem Gewebe oft schon belüftetes Lungenparenchym enthalten, was negative Auswirkungen auf die nachfolgenden Analysen und Radiomics-Features haben kann.¹⁷¹ Ein Maximaldurchmesser wurde nicht festgelegt, jedoch wurden mehrheitlich Rundherde <1cm im Durchmesser ausgewählt (siehe Ergebnisse).

Um eine mögliche Stichprobenverzerrung zu minimieren, wurde eine Obergrenze von maximal 10 Rundherden pro Patient*in gesetzt. Im Falle, dass mehr als 10 Herde die Einschlusskriterien erfüllten, wurde eine Auswahl von 10 geeigneten Herden getroffen, die eine größtmögliche morphologische Variabilität beinhalteten.

Diese Einschlusskriterien wurden bereits in einer früheren Radiomics-Studie des radiologischen Instituts der Uniklinik Köln verwendet. Diese Studie zeigte gute Klassifikationsergebnisse.⁶⁴ Um die Kohorte der vorliegenden Studie zu vergrößern, wurden die Rundherde dieser vorigen Studie (182 Patient*innen/Untersuchungen, darunter 150 benigne Rundherde und 408 Metastasen) im Verlauf in die Kohorte miteingeschlossen.

Eine Übersicht des in diesem und dem vorigen Abschnitt beschriebenen Selektionsprozesses ist in Abbildung 1 dargestellt.

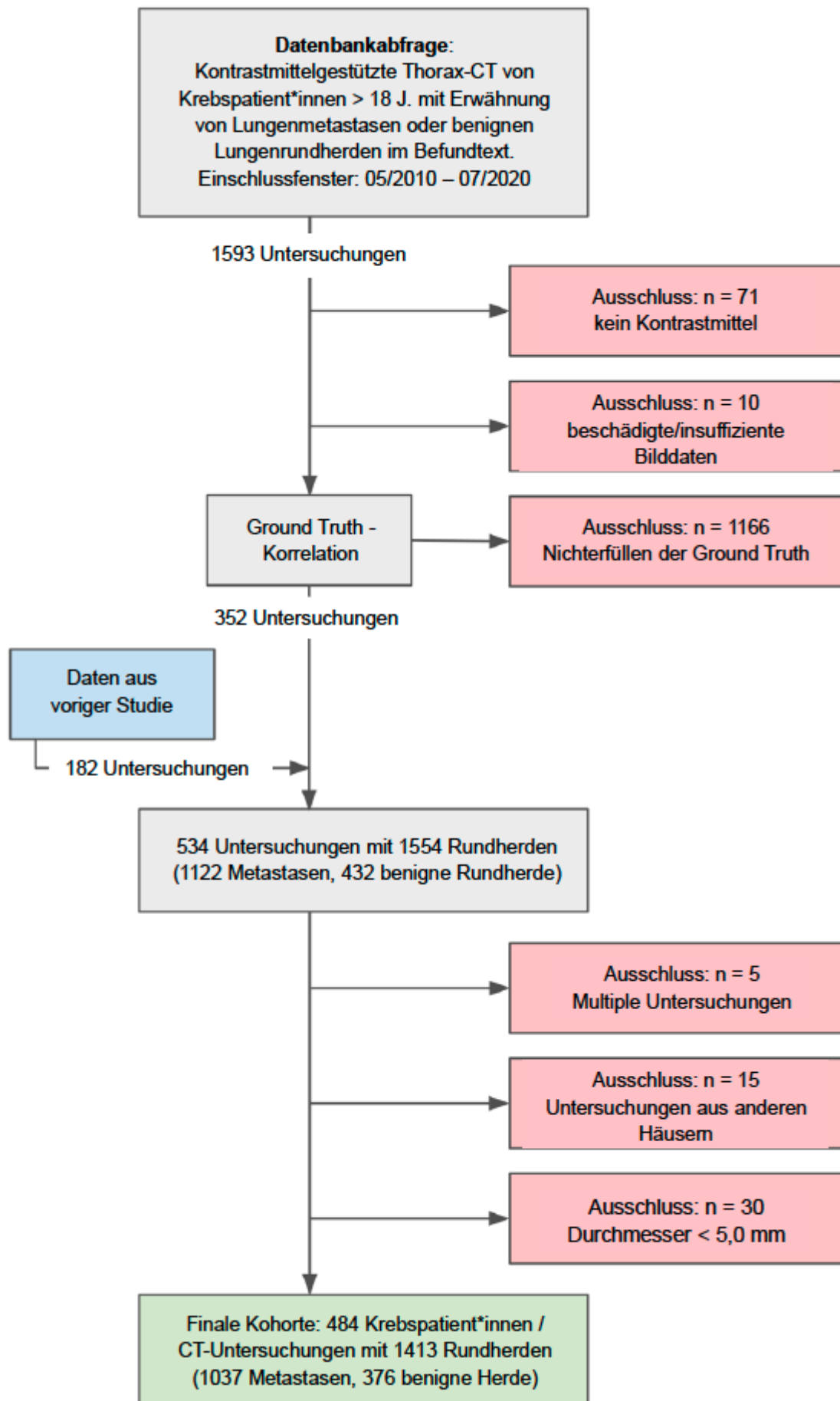


Abbildung 1: Selektion der Studienkohorte

3.2. Gewinnung und Rekonstruktion des Bildmaterials

Alle eingeschlossenen Untersuchungen erfolgten in einer Institution (Uniklinik Köln) unter standardisierten Protokollen für die kontrastmittelgestützte CT-Bildgebung des Thorax. Verschiedene CT-Geräte mehrerer Hersteller sowie unterschiedliche Rekonstruktionsalgorithmen, -kernel und Untersuchungsparameter wurden verwendet (siehe Tabelle 2).

Scanner-Typen

Hersteller	Scanner	Anzahl der eingeschlossenen Untersuchungen
Philips	IQon	268
Philips	iCT 256	109
Philips	Brilliance 64	37
Philips	Brilliance 6	24
Siemens	SOMATOM Force	30
Siemens	SOMATOM Definition Flash	16

Rekonstruktion

Hersteller	Rekonstruktions-Filter	
Philips	YA	173
Philips	L	139
Philips	IMR2, SharpPlus	107
Philips	YB	13
Philips	C	5
Philips	YC	1
Siemens	FLAT	28
Siemens	WEDGE_3	16
Siemens	WEDGE_2	2

Parameter

Schichtdicke	Schichtinkrement	
2.0	1.0	428
Weitere	Weitere	56
Matrix-Größe		
512	512	482
Weitere		2

3.3. Extraktion der Radiomics-Features aus Bilddaten

3.3.1. Semiautomatisierte Segmentierung der Lungenrundherde

Nach Auswahl der geeigneten Rundherde nach den unter 3.1.2. genannten Kriterien im Lungenfenster wurden die dazugehörigen Bilddateien im DICOM-Format in die Software mint Lesion research (mint medical GmbH, Heidelberg) übertragen. Zur Vorbereitung der Segmentierung erfolgte zunächst ein Training der hierfür eingesetzten Personen an einem separaten Datensatz von Rundherden, die nicht in die vorliegende Studie einbezogen wurden. Im nächsten Schritt erfolgte die semiautomatisierte Segmentierung der einzelnen Rundherde in allen Schichten. Hierbei wurden nur diejenigen Voxel markiert, die das Volumen der Rundherde repräsentierten. Peritumorales Lungenparenchym wurde nicht mitsegmentiert. Die Segmentierungen erfolgten in Rücksprache und unter Kontrolle eines Radiologen mit >5 Jahren Erfahrung in onkologischer Thoraxbildgebung. Bei unklaren Fällen wurde ein Radiologe mit >15 Jahren Erfahrung in diesem Feld konsultiert.

3.3.2. Interreader-Variabilitäts-Analyse

Wie bereits unter 3.1.2. beschrieben, wurde zur Vergrößerung der Studienkohorte ein zusätzlicher Datensatz in die Analyse einbezogen. Dieser umfasste letztendlich 558 segmentierte Rundherde von 178 Patient*innen aus einer zuvor durchgeführten Studie derselben Arbeitsgruppe, die identische Einschlusskriterien aufwies.⁶⁴ Zur Sicherstellung der Vergleichbarkeit und Wahrung der methodischen Integrität zwischen beiden Datensätzen wurden daher insgesamt 95 Rundherde des Datensatzes der vorigen Studie erneut durch den Auswerter der vorliegenden Studie segmentiert. Anschließend erfolgte ein Vergleich der Segmentierungsergebnisse beider Untersuchenden. Die Interreader-Variabilität wurde anhand des Dice-Scores (Verhältnis der zweifachen Anzahl gemeinsamer Voxel und der Gesamtsumme aller Voxel) und des Konkordanz-Korrelationskoeffizienten erfasst.

3.3.3. Extraktion der Radiomics-Features

Die Radiomics-Features wurden mithilfe des Programms PyRadiomics (Version 3.0.1) extrahiert bzw. berechnet.¹²⁵ Abgesehen von speziell für bestimmte Studien entwickelten Programmen ist PyRadiomics die am häufigsten eingesetzte Software zur Feature-Extraktion und verwendet die Definitionen der IBSI zur Standardisierung von Features.¹⁷⁴

Für die Feature-Extraktion wurden zunächst die Bild-und Segmentierungsdaten aus mint Lesion research exportiert. Die Daten wurden im Folgenden auf eine isotrope Voxelgröße ($1,0 \text{ mm}^3 \times 1,0 \text{ mm}^3 \times 1,0 \text{ mm}^3$) vereinheitlicht. Um fälschlich mitsegmentierte Knochen und

ausgedehnte Kalzifikationen sowie luftgefülltes Lungenparenchym aus den Analysen auszuschließen, wurden Voxel mit Hounsfield-Einheit-Werten >700 oder <-550 aus dem Datensatz entfernt.

Zur Extraktion der Graustufenhistogramm-basierten Features wurde eine Bin-Breite von 10 gewählt, um zu vermeiden, dass die Ergebnisse entweder zu detailliert oder zu stark gemittelt dargestellt werden (eng: undersmoothing/oversmoothing).¹⁷⁵

Insgesamt wurden 105 verschiedene Features aus den folgenden 6 Kategorien extrahiert: Shape, First Order, *grey level co-occurrence matrix* (glcm), *grey level dependence matrix* (gldm), *grey level run length matrix* (glrlm) und *neighborhood grey-tone difference matrix* (ngtdm). Zum Zweck der nachfolgenden Feature-Selektion wurden zusätzlich zu der gesamten Kohorte separat aus dem Datensatz der 95 Rundherde, die zur Analyse der Interreader-Variabilität verwendet wurden, alle 105 Features extrahiert, um anschließend deren Korrelation zu untersuchen (siehe 3.4.).

3.4. Statistische Analyse und Modellentwicklung

Eine schematische Darstellung des im Folgenden beschriebenen Arbeitsablaufes ist in Abbildung 1 dargestellt.

Zur Entwicklung des diagnostischen Klassifikationsmodells wurden das bereits unter 2.3.3. geschilderten Verfahren der verschachtelten Kreuzvalidierung angewandt. Der gesamte Datensatz aus insgesamt 1413 Rundherden wurde in 5 äußere Datensätze geteilt (äußere Schleife), welche jeweils im Verhältnis 80:20 in Datensätze zum Modelltraining (je 1130 Rundherde) und zur Modelltestung (je 283 Rundherde) unterteilt wurden. Diese Aufteilung erfolgte stratifiziert, um in Trainings- und Testanteil eine ähnliche Größenverteilung der Rundherde zu erhalten. Die äußeren Trainings-Datensätze wurden nun weiter in 5 innere Datensätze separiert (innere Schleife), erneut mit einem Trainings-Test-Verhältnis von 80:20 (je 904 Herde zum Training und 226 Herde zum Testen).

Vor Beginn der Modellentwicklung erfolgte eine Normierung der zuvor extrahierten Features mittels robuster Skalierung ($\text{skalierter Wert} = (\text{Originalwert} - \text{Median}) / \text{Interquartilabstand}$). Außerdem wurde die Methode des Random Undersampling auf die inneren Trainings-Datensätze angewandt, um ein bestehendes Klassenungleichgewicht durch randomisierte Löschung von Rundherden aus der überrepräsentierten Klasse zu beheben.

Innerhalb der inneren Trainings-Datensätze erfolgte daraufhin die Entwicklung der Klassifikationsmodelle, welche anschließend an den inneren Test-Datensätzen evaluiert wurden. Die Details der Modellentwicklung und -anwendung werden in den folgenden Kapiteln erklärt.

Zur finalen Bestimmung der Klassifikationsleistung wurden die fünf entwickelten Ensemble-Klassifikatoren auf die jeweiligen äußeren Testdatensätze angewandt. Für jeden Klassifikator wurden Sensitivität, Spezifität, AUC, F1-Score sowie Balanced Accuracy berechnet. Diese 5 Ergebnisse wurden abschließend gemittelt, um ein Gesamtergebnis abzubilden.

Für die im Folgenden beschriebenen Prozesse zur Modellentwicklung wurde die Programmiersprache Python (Version 3.8.13) mit den Packages pandas (Version 1.4.3)¹⁷⁰, NumPy (Version 1.22.3)¹⁷¹, scikit-learn (Version 1.0.2)¹⁷², XGBoost (Version 1.5.0)¹⁷³ und Precision-Medicine-Toolkit¹⁷⁶ verwendet. Der entsprechende Code wird in der Zukunft unter <https://github.com/datascienceukk/LungNodules> veröffentlicht (aktuell noch ungültiger Link, Stand 29.03.2026).

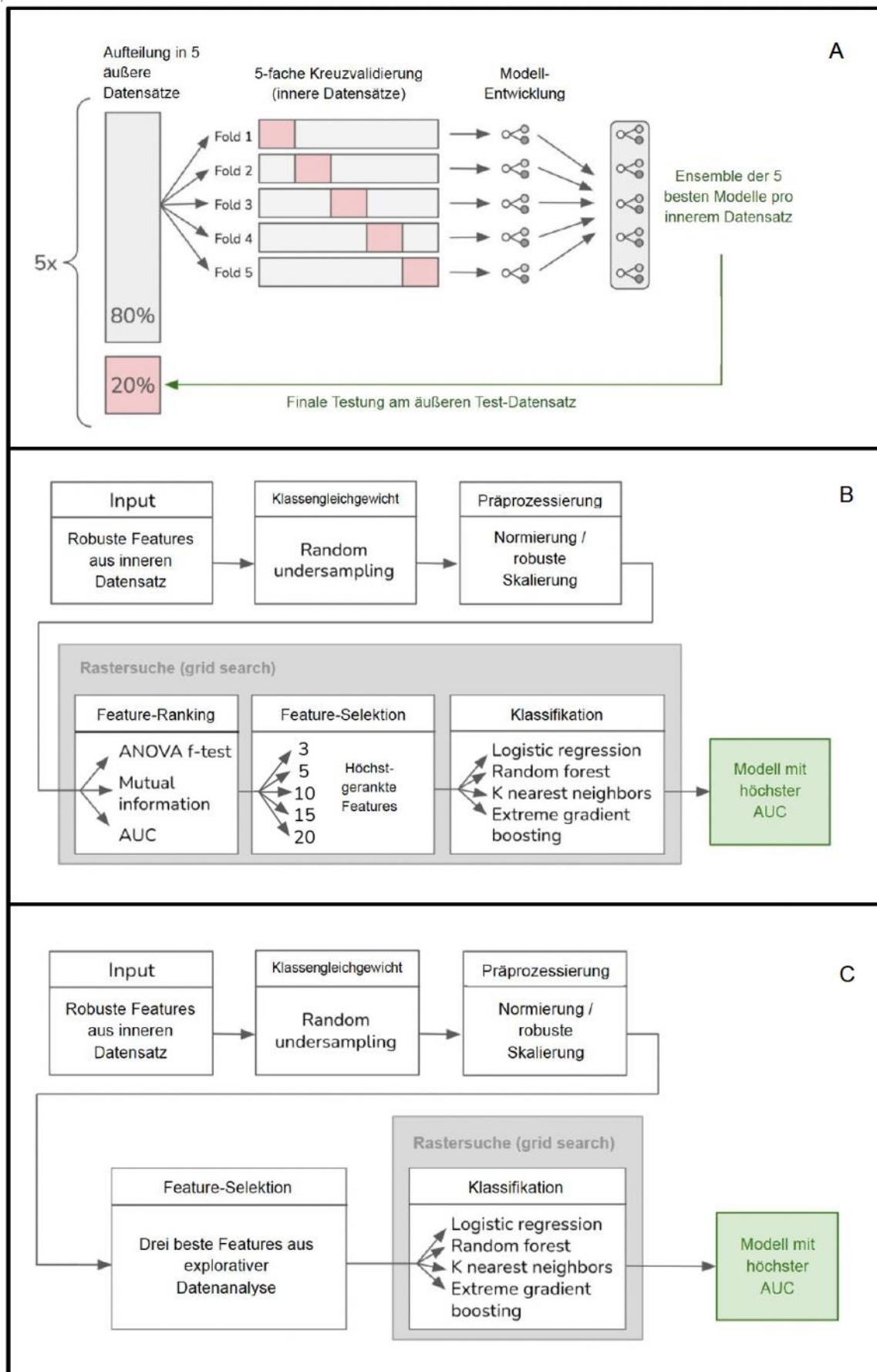


Abbildung 2: Übersicht des Arbeitsablaufes der Klassifikationsmodellerstellung. A: Übersicht der Modellentwicklungsprozesses unter Verwendung von Kreuzvalidierung. B: Detaillierte Darstellung der automatisierten Modellentwicklung. C: Detaillierte Darstellung der manuellen Modellentwicklung.

3.4.1. Entwicklung der Klassifikationsmodelle

Anstatt alle 105 extrahierten Features in Erwägung zu ziehen, wurden lediglich diejenigen Features berücksichtigt, die innerhalb des Datensatzes der Interreader-Variabilitätsanalyse Konkordanz-Korrelationskoeffizienten von $>0,75$ als Zeichen ihrer Robustheit aufwiesen.

Noch vor Beginn der Entwicklung eines Klassifikationsmodells wurden die im vorigen Schritt berechneten robusten Features mittels robuster Skalierung normiert (skalierter Wert = Originalwert - Median des Features / Interquartilabstand des Features). Außerdem wurde die Methode des Random Undersampling auf die inneren Trainings-Datensätze angewandt, um ein mögliches Klassenungleichgewicht durch randomisierte Löschung von Rundherden aus der überrepräsentierten Klasse zu beheben.

Für die weiteren Schritte der Modellentwicklung wurden im Rahmen dieser Studie zwei verschiedene Arbeitsabläufe angewandt, die aufgrund ihrer unterschiedlichen Methodik im Folgenden vereinfacht als automatisiert und manuell beschrieben werden.

Der automatisierte Arbeitsablauf (siehe Abbildung 1B) beinhaltet drei Komponenten: das Ranking der Features mittels der Metriken AUC-Score, ANOVA-f-Test und Mutual Information, die Selektion einer bestimmten Anzahl (3, 5, 10, 15 oder 20) an Features und die Klassifikation mittels logistischer Regression, Random Forest, K-Nearest Neighbors und XGBoost (eXtreme Gradient Boosting). Diese drei Komponenten wurden mittels einer Rastersuche (eng: grid search) zur Hyperparameter-Optimierung zu dem Modell mit der höchsten AUC für die vorliegende Klassifikationsfragestellung vereinigt, indem sämtliche Kombinationen systematisch getestet wurden.

Im manuellen Arbeitsablauf (Abbildung 1C) entfielen die automatisierten Prozesse des Feature-Rankings und der Feature-Selektion. Stattdessen wurde mithilfe des Python-Packages Precision-Medicine-Toolkit eine explorative Datenanalyse vorgenommen, die für jedes der robusten Features einzeln die Klassifikationsfähigkeit mittels AUC analysiert. Aus der Liste der Features, die $AUC > 0,7$ aufwiesen, wurden die Features ausgewählt, die eine geringe Korrelation (Spearman-Korrelation r $-0.50 \leq r \leq 0.50$) aufwiesen. Die anschließenden Klassifikationsmetriken zur finalen Modellerstellung blieben analog zum automatisierten Ablauf und wurden mit einer Rastersuche optimiert.

In jedem Datensatz der inneren Schleife wurde folglich ein Modell entwickelt. Diese wurden anschließend zu einem Ensemble-Klassifikator kombiniert. Dieser wurde anschließend am jeweiligen äußeren Testdatensatz angewandt und Sensitivität, Spezifität, ausgewogene Genauigkeit, Area Under the Receiver Operating Characteristic Curve (AUC) und F1-Score

berechnet. Insgesamt wurden entsprechend der fünf äußeren Datensätze fünf-Ensemble-Klassifikatoren entwickelt, deren Ergebnisse anschließend gemittelt wurden.

3.4.2. Analyse von Subgruppen

Neben des gesamten Datensatzes wurden auch spezielle Subgruppen des Datensatzes nach dem geschilderten Verfahren klassifiziert.

Wie bereits in 2.3.4. erwähnt, sind Radiomics-Features häufig nicht robust, wenn die CT-Bilddaten mit verschiedenen Algorithmen rekonstruiert wurden.¹³² Daher wurden Subgruppen für die häufigsten Rekonstruktionskernels gebildet und entsprechende Klassifikatoren entwickelt.

Neben der Bildrekonstruktionsverfahren wurden außerdem Subgruppen für die weiteren Untersuchung spezifischer Radiomics-Features gebildet. Zum einen wurde ein Klassifikator entwickelt, der nur auf Shape-Features basiert, zum anderen wurde ein weiterer Klassifikator ohne Berücksichtigung der Shape-Features entwickelt, um die Abhängigkeit der Klassifikationsfähigkeit von der Größe der Rundherde zu erfassen.

3.5. Evaluation der methodischen Qualität der Studie

Zur Evaluation der Qualität der vorliegenden Radiomics-Studie wurde der bereits unter 2.3.4. erwähnte Radiomics Quality Score (RQS)¹²⁸ erhoben. Der detaillierte Bewertungsbogen inklusive der Auswertung für die vorliegende Studie ist als Anhang 7.3. beigelegt.

4. Ergebnisse

4.1. Kohorte

Die Studienkohorte bestand aus 484 Patient*innen (213 weiblich, 271 männlich) mit insgesamt 1413 Rundherden. 1037 der Rundherde waren metastatischer Genese (davon 985 solide, 52 teilsolide) und 376 waren benigner Genese (davon 370 solide, 6 teilsolide). In der Kohorte mit Metastasen waren verschiedene onkologische Grunderkrankungen vertreten. Die häufigsten von ihnen waren das maligne Melanom und das Kolorektalkarzinom. Primärmalignome, deren Metastasen in weniger als fünf Patient*innen der Studienkohorte vertreten waren, wurden unter „Andere“ zusammengefasst. Diese sowie weitere demographische Charakteristika sind in Tabelle 3 dargestellt. Die Größe der Rundherde in den zwei verschiedenen Klassen ist in Tabelle 4 dargestellt.

Patient*innen	484
Alter	63,8 ± 13,1 Jahre
Geschlecht	271 männlich, 213 weiblich

Primärmalignome der Patient*innen mit Lungenmetastasen

Primärmalignom	Anzahl Patient*innen
Malignes Melanom	63
Andere	42
Kolorektalkarzinom	28
Mammakarzinom	20
Pankreaskarzinom	15
Nierenzellkarzinom	15
Lungenkarzinom	15
Ösophaguskarzinom	13
Sarkom, nicht näher bezeichnet	12
Urothelkarzinom	7
Ovarialkarzinom	6

Tabelle 3: Patient*innencharakteristika

	Metastasen	Benigne Rundherde
Volumen [mm³]	711,13 ± 9788,05	402,50 ± 2869,45
Größter axialer Durchmesser [mm]	9,35 ± 5,71	7,35 ± 5,39

Tabelle 4: Größe der Rundherde, berechnet durch die entsprechenden Shape-Features. Angaben in Mittelwert ± Standardabweichung.

4.2. Interreader-Variabilität

Um das Ausmaß der Variabilität zwischen den zwei Segmentierenden der Studie zu evaluieren, wurde durch direkten Vergleich der Segmentierungen an 95 Rundherden zunächst der Dice-Score berechnet. Dieser betrug $0,76 \pm 0,11$. Kleine Rundherde der Kohorte mit einem Volumen von $\leq 116,0 \text{ mm}^3$ zeigten teils geringe Dice-Scores von $<0,65$. Im Folgenden wurde außerdem der Konkordanz-Korrelationskoeffizient zwischen der Segmentationsvolumina berechnet. Dieser betrug 0,95. Die Ergebnisse der Interreader-Variabilitäts-Analyse sind in Tabelle 5 sowie Abbildung 3 dargestellt.

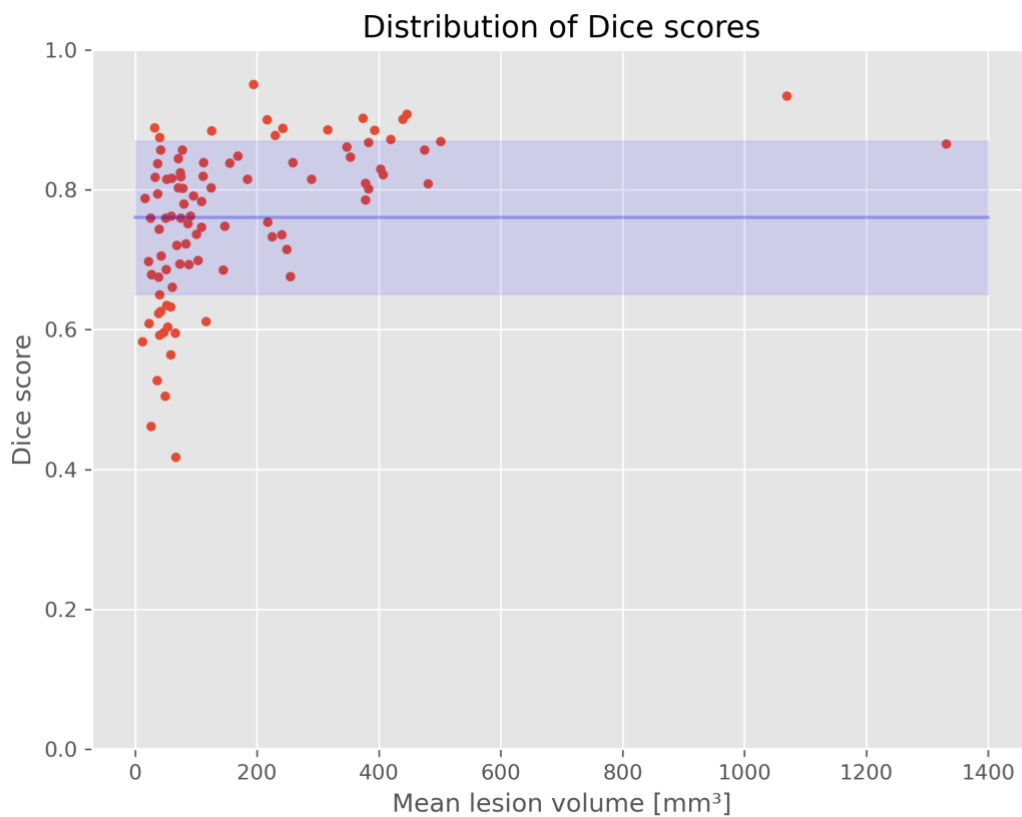


Abbildung 3: Interreader-Variabilitätsanalyse. Dice-Score in Relation zum Rundherd-Volumen.

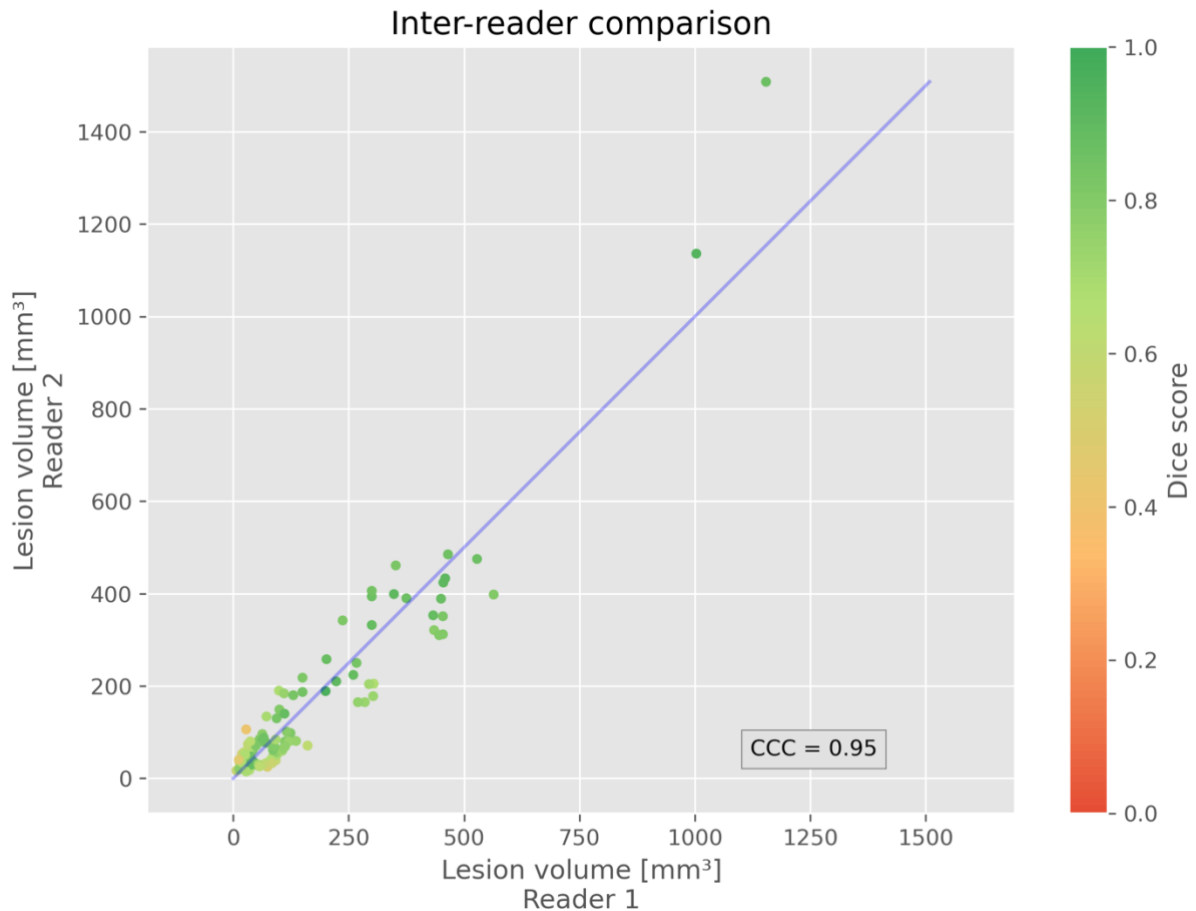


Abbildung 4: Interreader-Variabilitätsanalyse. Konkordanz-Korrelationskoeffizient in Relation zum Rundherd-Volumen

4.3. Entwicklung der Klassifikationsmodelle

Zur Identifizierung der robusten Features für die Klassifikationsmodelle wurden deren Korrelationskoeffizienten in dem Datensatz der Interreader-Variabilitätsanalyse berechnet. Von den 105 Features wurden sämtliche mit $CCC < 0,75$ ausgeschlossen (insgesamt 32). Dazu gehörten unter anderem *original_shape_Elongation*, *original_firstorder_Skewness* und *original_glcml_JointEnergy*. Eine Übersicht der einzelnen Features und deren CCC ist im Anhang beigefügt.

Im Rahmen der zuvor beschriebenen manuellen Feature-Selektion ergaben sich drei Features, die AUC-Werte $> 0,7$ sowie geringe Spearman-Korrelationen als Zeichen der Unabhängigkeit aufwiesen: *original_shape_Flatness* (AUC $0,73 \pm 0,01$), *original_shape_SurfaceVolumeRatio* (AUC $0,79 \pm 0,00$), und *original_glcml_ClusterShade* (AUC $0,72 \pm 0,01$). Die Spearman-Korrelationen betrugen $-0,42 \pm 0,01$ für *original_shape_Flatness* und *original_shape_SurfaceVolumeRatio*, $0,48 \pm 0,01$ für *original_glcml_ClusterShade* und *original_shape_SurfaceVolumeRatio* sowie $-0,25 \pm 0,01$ für *original_shape_Flatness* und *original_glcml_ClusterShade*.

4.4. Klassifikationsergebnisse

Eine Übersicht über die Klassifikationsergebnisse der Ensemble-Klassifikatoren, die in den zuvor beschriebenen zwei unterschiedlichen Arbeitsabläufen entwickelt wurden, ist in den Tabellen 5 und 6 dargestellt.

<i>Subgruppe/Metrik</i>	Sensitivität	Spezifität	F1-Score	Area Under Curve (AUC)	Balanced Accuracy
Alle Untersuchungen und Features	0,72 ± 0,07	0,75 ± 0,02	0,60 ± 0,04	0,81 ± 0,03	0,74 ± 0,03
IMR Sharp Plus	0,77 ± 0,12	0,75 ± 0,08	0,61 ± 0,08	0,84 ± 0,03	0,76 ± 0,06
iDose YA	0,76 ± 0,07	0,75 ± 0,08	0,69 ± 0,03	0,81 ± 0,02	0,75 ± 0,03
iDose L	0,65 ± 0,10	0,72 ± 0,10	0,46 ± 0,07	0,77 ± 0,06	0,68 ± 0,07
Shape-Features	0,71 ± 0,08	0,74 ± 0,03	0,58 ± 0,03	0,80 ± 0,02	0,73 ± 0,03
Ausschluss von Shape-Features	0,71 ± 0,07	0,72 ± 0,03	0,57 ± 0,02	0,78 ± 0,03	0,72 ± 0,02

Tabelle 5: Klassifikationsergebnisse des Ensemble-Klassifikators unter Verwendung des automatisierten Arbeitsablaufes zur Klassifikatorentwicklung. Angaben in Mittelwert ± Standardabweichung.

<i>Subgruppe/Metrik</i>	Sensitivität	Spezifität	F1-Score	Area Under Curve (AUC)	Balanced Accuracy
Alle Untersuchungen und Features	0,78 ± 0,11	0,68 ± 0,13	0,59 ± 0,05	0,82 ± 0,03	0,73 ± 0,05
IMR Sharp Plus	0,80 ± 0,09	0,70 ± 0,15	0,60 ± 0,07	0,83 ± 0,04	0,75 ± 0,04
iDose YA	0,75 ± 0,04	0,76 ± 0,05	0,69 ± 0,01	0,81 ± 0,02	0,76 ± 0,01
iDose L	0,79 ± 0,09	0,72 ± 0,06	0,53 ± 0,02	0,79 ± 0,04	0,75 ± 0,03

Tabelle 6: Klassifikationsergebnisse des Ensemble-Klassifikators unter Verwendung des manuellen Arbeitsablaufes zur Klassifikatorentwicklung. Aufgrund der manuell selektierten Features ist keine Testung der Feature-Subkategorien möglich. Angaben in Mittelwert ± Standardabweichung

4.4.1. Gesamte Kohorte

Unter Verwendung der automatisierten Modellentwicklung konnte für die Klassifikation der beginnenden Rundherde in der gesamten Kohorte eine mittlere Sensitivität von 0,72 ± 0,07 und eine mittlere Spezifität von 0,75 ± 0,02 erreicht werden. Die AUC betrug 0,81 ± 0,03. Die Klassifikationsmodelle, die mit manuell selektierten Features entwickelt wurden, erreichten ähnliche Ergebnisse (Sensitivität 0,78 ± 0,11, Spezifität 0,68 ± 0,13, AUC 0,82 ± 0,03). Die Ergebnisse der Klassifikation basierend auf manuell und automatische extrahierten Features sind in den Abbildungen 5 und 6 dargestellt.

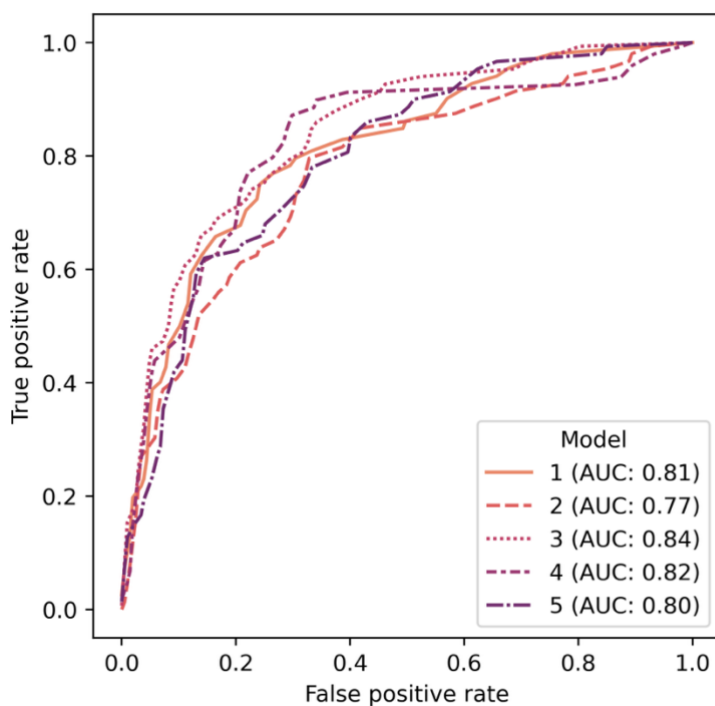


Abbildung 5: Klassifikationsergebnisse innerhalb der gesamten Studienkohorte unter Verwendung der automatisierten Feature-Selektion. ROC-Kurve der einzelnen äußeren Test-Datenätze und Darstellung der Falsch-Positiv-Rate (1-Spezifität) und Echt-Positiv-Rate (Sensitivität)

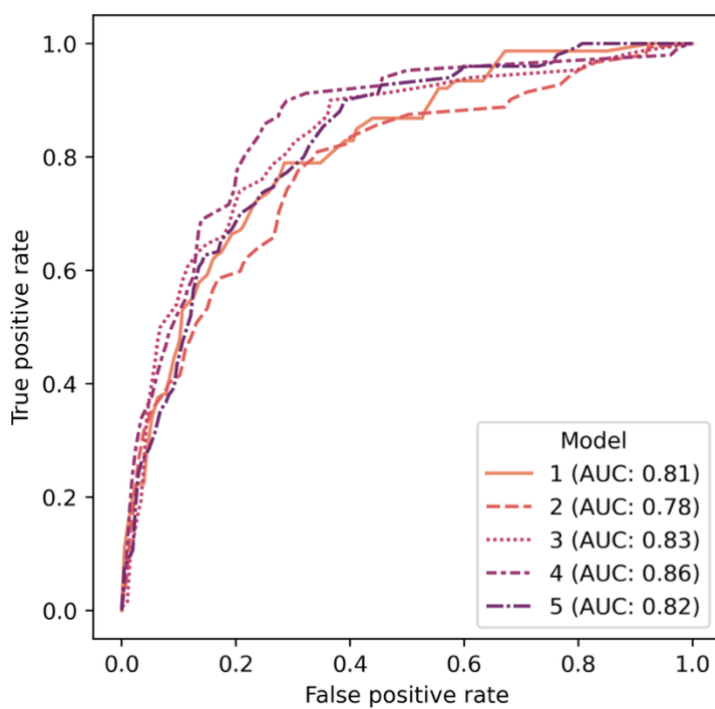


Abbildung 6: Klassifikationsergebnisse innerhalb der gesamten Studienkohorte unter Verwendung der manuell extrahierten Features nach explorativer Datenanalyse. ROC-Kurve der einzelnen äußeren Test-Datenätze und Darstellung der Falsch-Positiv-Rate (1-Spezifität) und Echt-Positiv-Rate (Sensitivität).

4.4.2. Rekonstruktionen

Zur Untersuchung des Einflusses der verschiedenen Bildrekonstruktionsverfahren (siehe Tabelle 2) wurden entsprechende Subgruppen analysiert. Drei dieser Gruppen waren ausreichend groß für eine statistische Analyse: Die IMR-Rekonstruktion mit dem Kernel Sharp Plus als CT-Lungenfenster (307 Rundherden, davon 78 benigne Rundherde und 229 Metastasen) sowie die iDose-Rekonstruktion mit den Filtern YA (449 Rundherde, davon 160 benigne Rundherde und 289 Metastasen) und L (421 Rundherde, davon 80 benigne Rundherde und 341 Metastasen). Es wurden spezielle Klassifikatoren trainiert, die einzig die Daten der Subgruppen verwendeten.

Hier zeigten sich variierende Ergebnisse mit Sensitivitäten zwischen $0,65 \pm 0,10$ (iDose L) und $0,77 \pm 0,12$ (IMR Sharp Plus), Spezifitäten zwischen $0,72 \pm 0,10$ (iDose L) und $0,75 \pm 0,08$ (IMR Sharp Plus und iDose YA). Die höchste AUC wurde mittels IMR Sharp Plus erreicht ($0,84 \pm 0,03$). Hervorzuheben sind zudem deutliche Ergebnisunterschiede im F1-Score (iDose L: $0,46 \pm 0,07$, iDose YA: $0,69 \pm 0,03$).

Die gesammelten Ergebnisse sind in den Tabellen 5 und 6 dargestellt.

4.4.3. Feature-Subgruppen

Zwei der drei im Rahmen der manuellen Feature-Selektion gewählten Features gehörten der Kategorie der Shape-Features an (*original_shape_Flatness* und *original_shape_SurfaceVolumeRatio*). Aus diesem Grund wurden im Verlauf im Rahmen der automatisierten Feature-Selektion zwei zusätzliche Klassifikatoren entwickelt: Einer, der ausschließlich Shape-Features enthielt und ein weiterer, der Shape-Features explizit ausschloss. Ersterer zeigte eine Sensitivität von $0,71 \pm 0,08$, eine Spezifität von $0,74 \pm 0,03$ und eine AUC von $0,80 \pm 0,02$. Letzterer zeigte Ergebnisse, die innerhalb der Standardabweichung des ersteren lagen (Sensitivität $0,71 \pm 0,07$, Spezifität $0,72 \pm 0,03$, AUC $0,78 \pm 0,03$). Die Ergebnisse sind in Tabelle 5 dargestellt.

4.5. Evaluation der methodischen Qualität der Studie

Zur Evaluation der Qualität der Studie wurde der von Lambin et al. entwickelten Radiomics Quality Score (RQS)¹²⁸ verwendet. Dieser ergab eine Gesamtpunktzahl von 11 Punkten (30,56 %). Der vollständig ausgefüllte Bewertungsbogen ist unter Anhang 7.3. beigefügt.

5. Diskussion

Die vorliegende Studie widmete sich der Differenzierung benigner Lungenrundherde von Lungenmetastasen in onkologischen Patient*innen mittels eines CT-Radiomics-basierten Klassifikationsmodells. Diese Fragestellung adressiert eine bisherige diagnostische Herausforderung in der Thoraxbildgebung, da für die sichere Charakterisierung von Lungenrundherden innerhalb einer Einzeluntersuchung bei onkologischen Patient*innen keine etablierten oder diagnostischen Kriterien bestehen, ein korrektes M-Staging jedoch für eine adäquate Therapieentscheidung essenziell ist. Ziel der Studie war es außerdem, ein Modell zu entwickeln, das gut über mehrere CT-Geräte sowie verschiedene onkologische Grunderkrankungen generalisierbar ist, um eine möglichst breite, praxisnahe Anwendbarkeit zu ermöglichen.

Mithilfe der entwickelten Klassifikationsmodelle konnten eine mittlere Sensitivität von $0,72 \pm 0,07$, eine mittlere Spezifität von $0,75 \pm 0,02$ und eine AUC von $0,81 \pm 0,03$ erreicht werden. Der Einfluss verschiedener Rekonstruktionsalgorithmen und einzelner Radiomics-Features wurden ebenfalls untersucht. Zudem wurden zwei verschiedene Methoden der Modellerstellung verwendet. Diese Studie ist nach jetzigem Kenntnisstand des Autors die erste, die diese klinisch relevante Klassifikationsaufgabe mit kontrastmittelgestützten CT-Daten in einer derart heterogenen, großen Kohorte untersucht hat.

Im Folgenden werden Studiendesign und -ergebnisse in ihren wissenschaftlichen Kontext eingeordnet, Limitationen aufgezeigt und Empfehlungen für kommende Forschungsprojekte ausgesprochen.

5.1. Studienkohorte

Die vorliegende Studienkohorte bestand aus 484 Patient*innen mit insgesamt 1413 Lungenrundherden. Maschinelle Lernprozesse profitieren von großen Datensätzen, um generalisierbare Ergebnisse, z.B. eine leistungsstarkes Klassifikationsmodell, zu erreichen.¹⁷⁷ Grundsätzlich wird empfohlen, etwa das Zehnfache an Datenpunkten, in diesem Fall Rundherde, im Verhältnis zur Anzahl der Features zu haben.¹⁷⁸ In der vorliegenden Untersuchung wurden pro Modell maximal 20 Features selektiert, während im Trainingsdatensatz über 900 Rundherde enthalten waren, sodass diese Anforderung erfüllt wurde.

Die Geschlechterverteilung unserer Kohorte lag bei 44 % Frauen (213 Rundherde) und 56 % Männern (271 Rundherde). Diese nahezu ausgeglichene Verteilung trägt zur Generalisierbarkeit unserer Ergebnisse bei.

Im Gegensatz zu Radiomics-Studien mit ausgeglichenen Kohorten, z.B. Beig et al.¹⁶⁶, zeigte die Kohorte dieser Studie ein deutliches Klassenungleichgewicht: 1037 (73 %) der Rundherde waren metastatischer und 376 (26 %) waren benigner Genese. Dies birgt das Risiko eines Overfittings. Wie bereits unter 2.3.3. erwähnt, beschreibt dieses Phänomen eine während des Trainings entstehende Bias des Klassifikationsmodells zur überrepräsentierten Klasse des Datensatzes. Dies kann folglich zu falschen Klassifikationen führen, was im Kontext onkologischer Diagnostik weitreichende Konsequenzen haben kann.¹¹⁸ Um einer solchen Bias entgegenzuwirken, gibt es verschiedene Ansätze zur Reduktion eines Klassenungleichgewichts.

In dieser Studie wurde Random Undersampling angewandt. Dies beinhaltet eine randomisierte Löschung von Datenpunkten aus der überrepräsentierten Klasse.¹³⁹ Diese Methode ist aufgrund der Verkleinerung des Datensatzes schnell und kostengünstig¹⁷⁹, birgt jedoch durch den konsekutiven Informationsverlust das Risiko, potenziell klassifikationsrelevante Muster zu verbergen.^{180,181} Andere Methoden wie zum Beispiel SMOTE (Synthetic Minority Over-Sampling Technique)¹⁸² und deren Weiterentwicklungen wurden in der Vergangenheit als effektiver beschrieben.^{179,183} Ein Vergleich mit einer Subgruppe der Kohorte, in der kein Random Undersampling oder eine andere Methoden wie SMOTE angewandt wurde, um deren Effekt einzuschätzen, wurde jedoch nicht vorgenommen. In Zukunft sollten im Falle eines Ungleichgewichts auch weitere Methoden berücksichtigt werden. Generell sollte jedoch eine ausgeglichene Kohorte das Ziel sein.

Die Analyse der Herdgrößen ergab, dass benigne Rundherde mit einem durchschnittlichen Volumen von $402,50 \pm 2869,45 \text{ mm}^3$ und einem mittleren Durchmesser von $7,35 \pm 5,39 \text{ mm}$ kleiner waren als Metastasen, die im Mittel ein Volumen von $711,13 \pm 9788,05 \text{ mm}^3$ und einen Durchmesser von $9,35 \pm 5,71 \text{ mm}$ aufwiesen. Zum Zeitpunkt der Datenerhebung der vorliegenden Studie galten die Fleischner-Definitionen von 2008, in denen der Minstdurchmesser eines Rundherds als $\geq 3 \text{ mm}$ festgelegt wurde.¹⁸⁴ Auf Basis dieser Definition wurde für diese Studie ein Minstdurchmesser von 5 mm als Einschlusskriterium festgelegt. Die Segmentierung der Rundherde unterhalb dieser Schwelle ist häufig schwierig, da einzelne Voxel im Randbereich der Rundherde neben relevanten Rundherdanteilen oft schon belüftetes Lungenparenchym enthalten, was negative Auswirkungen auf die nachfolgenden Analysen und Radiomics-Features haben kann.¹⁷¹ Hu et al. verzichteten in ihrer Studie aus diesen Gründen auf die Segmentierung des Randbereichs.¹⁷² Ein

Maximaldurchmesser wurde nicht festgelegt, jedoch wurden mehrheitlich Rundherde <1 cm im Durchmesser ausgewählt (siehe Tabelle 4). Dieser Fokus beruhte darauf, dass insbesondere kleinere Rundherde eine große diagnostische Herausforderung darstellen, u.a. da sie schwierig zu biopsieren sind¹⁸⁵ und anfällig für falsch-negative Ergebnisse in der PET-CT sind.⁴⁴ Im Rahmen der Suche nach Rundherden, die die Einschlusskriterien erfüllten, musste eine Vielzahl kleinerer, benigner Rundherde aufgrund des Minstdurchmessers von 5 mm ausgeschlossen werden. Metastatische Rundherde dagegen waren häufig größer und grundsätzlich zahlreicher für die Studie zu gewinnen. Dies trug zum beschriebenen Klassenungleichgewicht bei. Entsprechend unserer Einschlusskriterien konnten bis zu zehn Rundherde pro Patient*in die Kohorte aufgenommen werden. Dies war jedoch fast ausschließlich bei metastatischen Rundherden möglich. Ließen sich mehr als zehn geeignete Herde identifizieren, wurden die überschüssigen Rundherde von der Analyse ausgeschlossen. Die Auswahl der Metastasen erfolgte hierbei nicht randomisiert, weshalb eine mögliche Verzerrung der Studienkohorte nicht ausgeschlossen werden kann, sondern mit dem Ziel, Rundherde mit einer größtmöglichen morphologischen Variabilität in die Studie einzuschließen.

Ein weiteres methodisches Merkmal dieser Arbeit ist die Einbeziehung von Metastasen verschiedener Primärtumoren (siehe Tabelle 3). Ziel der Studie war es, ein Klassifikationsmodell zu entwickeln, das eine hohe Generalisierbarkeit aufweist. Bisherige Radiomics-Studien zur Klassifikation von unklaren Läsionen untersuchten mehrheitlich nur Metastasen einzelner Krebsarten.^{52,81,172} Jansen et al. dagegen unterschieden fünf verschiedene Arten von Leberläsionen mittels MRT-Radiomics. Hier wurden Metastasen verschiedener Grunderkrankungen zu einer Klasse kombiniert. Aufgrund der morphologischen Unterschiede verschiedener Metastasen und der daraus resultierenden Variabilität der Features wurden für die Klasse der Metastasen die niedrigsten Sensitivitäts-, Genauigkeits- und AUC-Werte aller fünf Typen an Leberläsionen erreicht. Gleichzeitig konnte die Mehrheit der Metastasen trotzdem korrekt klassifiziert werden (AUC = 0,88) und die drei häufigsten Metastasentypen zeigten kaum Unterschiede in der Genauigkeit der Klassifikation.¹⁸⁶ Gatos et al. untersuchten ebenfalls Lebermetastasen mit MRT-Radiomics und zeigten für die Klasse der Metastasen sogar leicht höhere Werte für Genauigkeit, Sensitivität und Spezifität als die Klasse der hepatozellulären Karzinome.¹⁸⁷ Lennartz et al. unterschieden Lungenmetastasen mehrerer Grunderkrankungen und benigne Lungenrundherde mittels CT-Radiomics und erreichten Genauigkeitswerte von 0,87.⁶⁴

Vor diesem Hintergrund erscheint die in dieser Studie gewählte Vorgehensweise der Zusammenfassung verschiedener Metastasenarten zu einer Klasse gerechtfertigt bzw. von klinischer Perspektive vielmehr gerade erwünscht.

Die vorliegende Studie untersuchte sowohl solide als auch subsolide Rundherde. Subsolide Rundherde gewinnen in der klinischen Praxis an Bedeutung, da sie zunehmend häufiger diagnostiziert werden¹⁸⁸ und sowohl mit Primärmalignomen als auch Metastasen der Lunge assoziiert sind.¹⁸⁹ Darüber hinaus zeigen sie in der Regel ein langsames Wachstum und eine geringere Stoffwechselaktivität als solide Rundherde.¹⁹⁰ Dementsprechend sind bereits Studien erfolgt, die explizit subsolide Rundherde untersuchen.^{191,192} Die Klassifikationsergebnisse dieser Studie legen jedoch nahe, dass das entwickelte Modell auf Rundherde aller Dichtegrade übertragbar ist.

5.2. Inter- und Intrareader-Variabilität

Die Segmentierungen der 1413 Rundherde wurden von zwei verschiedenen Personen durchgeführt. Dies war darin begründet, dass zur Vergrößerung der Kohorte Rundherde einer früheren Radiomics-Studie dieser Arbeitsgruppe mit gleichen Einschlusskriterien⁶⁴ in diese Studie hinzugefügt wurden. Die Interreader-Variabilität ist ein wichtiger Faktor, der in diesen Fällen berücksichtigt werden muss. Unterschiede in der Segmentierung zwischen den beiden Auswertern können die Radiomics-Features und damit auch die Leistung der entwickelten Klassifikationsmodelle beeinflussen.¹⁰⁵ Aus diesem Grund wurden zur Sicherung der Integrität des gesamten Datensatzes 95 der Rundherde von beiden Readern segmentiert und die Interreader-Variabilität mittels des Dice-Scores und des Konkordanz-Korrelationskoeffizienten untersucht.

Der mittlere Dice Score betrug $0,76 \pm 0,11$. Da dieser Score niedriger als erwartet war, wurde der Zusammenhang zwischen Dice Score und Rundherdvolumen untersucht. Der Dice-Score vergleicht ausschließlich einzelne Pixel bzw. Voxel.¹⁹³ Wie bereits in Tabelle 4 dargestellt betrug der durchschnittliche Durchmesser der eingeschlossenen Herde <10 mm. Da kleine Rundherde zum Teil nur aus wenigen Voxel bestehen, ist anzunehmen, dass Unterschiede von nur einem Voxel Metriken wie den Dice-Score und die eventuellen Features folglich stark beeinflussen können.¹³³ Es zeigte sich, dass Dice-Scores von $<0,65$ ausschließlich bei Rundherden mit einem Volumen von $\leq 116,0$ mm³ auftraten. Dies trifft mehrheitlich auf die kleinsten Herde zu, die analysiert wurden. Das durchschnittliche Volumen der eingeschlossenen Metastasen betrug 711,13 mm³ und das der benignen Rundherde 402,50 mm³, ein Vielfaches von 116,0 mm³. Daraus wurde geschlossen, dass der Dice-Score bei derartig kleinen Rundherden weniger Aussagekraft hat. Aus diesem Grund wurde zusätzlich der CCC als Maß der Interreader-Variabilität berechnet. Dieser lag bei 0,95, was einer geringen Variabilität der extrahierten Features entspricht. Aus diesem Grund wird der Einfluss der Interreader-Variabilität auf die Ergebnisse der Studie als gering eingeschätzt.

Neben der Interreader-Variabilität ist auch die Intrareader-Variabilität ein möglicher Einflussfaktor auf die Robustheit und Reproduzierbarkeit der Radiomics-Features und damit auf die Klassifikatoren.¹⁰⁰ Eine Untersuchung dieser Variabilität erfordert die Durchführung der Rundherdsegmentierungen an zwei verschiedenen Zeitpunkten, gefolgt von einem Vergleich der resultierenden Features.¹⁹⁴ Eine solche Analyse wurde nicht durchgeführt, da sie außerhalb des Studienziels lag.

Um den möglichen Einfluss von Intra-und Interreader-Variabilität gänzlich zu umgehen, kommen vollautomatisierte Segmentierungstechniken infrage. Jedoch besteht auch hier das Risiko einer mangelnden Generalisierbarkeit der entsprechenden Algorithmen.¹⁰⁰ Daher wurde in dieser Studie auf diese Form der Segmentierung verzichtet.

5.3. Klassifikationsergebnisse

5.3.1. Gesamte Kohorte

Für die gesamte Patient*innenkohorte ungeachtet der verschiedenen Feature-Kategorien und Rekonstruktionsalgorithmen konnte die finale Ensemble-Klassifikator Sensitivitäten von $0,72 \pm 0,07$ und $0,78 \pm 0,11$ sowie Spezifitäten von $0,75 \pm 0,01$ und $0,68 \pm 0,13$ für die Identifikation benigner Rundherde in einer Kohorte aus benignen Rundherden und Lungenmetastasen erreichen. Die AUC betrugen $0,81 \pm 0,03$ und $0,82 \pm 0,03$.

Studien, die explizit die Unterscheidung benigner und metastatischer Rundherde untersuchen, sind weiterhin selten, während die Unterscheidung zwischen benignen Rundherden und Lungenkarzinomen breit erforscht ist.⁶⁴

In einer vorigen Studie des radiologischen Instituts der Uniklinik Köln entwickelten Lennartz et al. ein Klassifikationsmodell für benigne und metastatische Rundherde verschiedener onkologischer Grunderkrankungen und konnten eine maximale Balanced Accuracy von 0,87 erreichen.⁶⁴ Während diese Studie identische Einschlusskriterien wie die aktuelle verwendete und die Studienkohorte in diejenige der aktuellen Studie miteinbezogen wurde, bestehen jedoch mehrere methodische Unterschiede: Die Studie von Lennartz et al. verwendete ausschließlich ein Dual-Energy-CT, um neben konventionellen CT-Aufnahmen auch Iodkarten generieren zu können. Zwar konnte die Kombination beider Aufnahmen die höchste Balanced Accuracy erreichen, jedoch zeigte sich im Vergleich zu dem Modell aus rein konventionellen CT-Features nur eine geringfügig bessere Genauigkeit (0,86 gegen 0,84). Aus diesem Grund wurde in der aktuellen Studie auf DECT verzichtet, um eine größere, heterogenere Studienpopulation gewinnen zu können. Außerdem wurden in der vorigen Studie die Metastasen als positive Klasse der Klassifikationsfrage gewählt. Da die erreichten Ergebnisse jedoch nicht ausreichend gewertet wurden, um bei hochsuspekten Rundherden invasive

Biopsien mit potenziell weitreichenden therapeutischen Konsequenzen vollständig zu ersetzen, wurden in der vorliegenden Arbeit die benignen Rundherde als positive Klasse gewählt. Eine verlässliche Diagnose benigner Herde durch Radiomics könnte dagegen unnötige Biopsien mit Komplikationsrisiken verringern.

Weitere Studien untersuchen vorrangig Lungenmetastasen einer bestimmten Tumorentität und/oder befassen sich ausschließlich mit der Unterscheidung von Metastasen und Nicht-Metastasen. Letztere Gruppe beinhaltet z.T. auch primäre Lungenkarzinome oder entsprechende maligne Vorstufen. Diese Studien sind daher nur begrenzt mit der vorliegenden Studie zu vergleichen, in der Primärmalignome gänzlich ausgeschlossen wurden. Dies verdeutlicht noch einmal die diagnostische Problematik, für die in dieser Studie Lösungsansätze gesucht wurden.

Cho et al. unterschieden pulmonale Osteosarkom-Metastasen von nicht-metastatischen Läsionen mittels Texture-Features und erreichten eine AUC von 0,865.¹⁷¹ Anders als in unserer Studie wurden bei Cho et al. rein pädiatrische Patient*innen eingeschlossen. Die Gruppe der Nicht-Metastasen beinhaltete einen Fall einer atypischen adenomatösen Hyperplasie. Letztere ist eine Vorläuferläsion des invasiven Adenokarzinoms⁵⁹ und gilt daher nicht als rein benigne Läsion. Außerdem war die Kohorte mit 24 Metastasen und 18 Nicht-Metastasen sehr klein und das entwickelte Klassifikationsmodell wurde nicht an unabhängigen Rundherden validiert, die an der Modellerstellung nicht beteiligt waren, was wiederum das Risiko eines Overfittings birgt. Einer der Gründe für die derart kleine Kohorte war das Einschlusskriterium einer histopathologischen Bestätigung der Dignität der untersuchten Rundherde, die den Goldstandard der Rundherd-Diagnostik darstellt.¹⁹⁵ Die vorliegende Studie verzichtete auf die Histopathologie als verpflichtendes Einschlusskriterium zugunsten des Wachstumsverhaltens der Rundherde in Folgeuntersuchungen und des malignomtypischen Radionuklid-Uptakes in der PET-CT, welche neben der Histopathologie als Ground Truth akzeptiert wurden. Dies erlaubte die Rekrutierung einer größtmöglichen Kohorte. Gleichzeitig besteht ohne histopathologische Diagnose jedoch das Risiko einer fehlerhaften Annotierung der Rundherde: Lungenmetastasen zeigen z.T. ein langsames Wachstum mit Tumorverdopplungszeiten von bis zu einem Jahr⁷³, subsolider wachsende Karzinome sogar von mehreren Jahren.^{50,51} Es besteht folglich die Möglichkeit, dass solch langsam wachsende Malignome oder auch Tumor-/Metastasenresiduen aus diesem Grund im Rahmen der Rekrutierung der Kohorte nach den unter 3.1.2 genannten Kriterien fälschlicherweise als benigne Rundherde klassifiziert werden und die aus ihnen gewonnenen Features den Klassifikator negativ beeinflussen konnten.⁶⁴ Im Rahmen dieser Studie wurde insbesondere bei solch uneindeutigen Herden nach Möglichkeit die gesamte Bildgebungshistorie der Patient*innen über die in den Einschlusskriterien formulierten sechs

Monate hinaus in die Evaluation miteinbezogen und Rücksprache mit mehreren Radiologen gehalten.

Die Rekrutierung einer Kohorte der Größe dieser Studie mit durchweg histopathologisch bestätigten Lungenrundherden, die außerdem die unter 3.1.1. und 3.1.2. erläuterten Einschlusskriterien erfüllen, ist empfehlenswert, wäre jedoch mit deutlich größerem Aufwand und der Zusammenarbeit mehrerer Kliniken verbunden, insbesondere wenn ein Gleichgewicht benigner und metastatischer Lungenrundherde angestrebt werden würde. Ein multizentrischer Ansatz birgt zudem aufgrund standortspezifischer CT-Parameter das Risiko einer geringen Robustheit und Reproduzierbarkeit der Features.

Die Studie von Cho et al. verwendete im Gegensatz zur vorliegenden Studie eine manuelle Segmentierung der Rundherde.¹⁷¹ Dies stellt eine zeitintensive Methode dar und ist im Gegensatz zu semi- und vollautomatisierten Segmentierungen anfällig für die Störfaktoren hoher Inter- und Intra-reader-Variabilitäten, was zu nicht robusten oder reproduzierbaren Features führen kann.¹⁰⁰ Die Korrelationskoeffizienten fielen in der Studie jedoch sehr hoch aus, was für eine sorgfältige Arbeit spricht.¹⁷¹

Hu et al. führten zwei Studien zur Klassifikation von Lungenmetastasen von Kolorektalkarzinomen durch.^{172,173} In beiden Studien wurden wie bei Cho et al. (s.o.) benigne Rundherde und Primärmalignome der Lunge zu einer Klasse zusammengefasst. Die Validierungskohorten der beiden Studien zeigten höhere AUC-Werte als unsere Studie: 0,929¹⁷² und 0,895.¹⁷³

Die entwickelten Klassifikationsmodelle verwendeten neben eines Radiomics-Scores weitere klinische und radiomorphologische Marker, bspw. Lymphknotenstatus und Rundherd-Durchmesser.¹⁷² Derartige Kombinationsmodelle zeigen häufig genauere Ergebnisse als die einzelnen, isolierten Komponenten: Ding et al. kombinierten die Variablen des Patient*innenalters, der Serumkonzentrationen sieben verschiedener Antikörper und des Carcinoembryonalen Antigens (CEA) mit aus CT extrahierten Features zu einem diagnostischen Nomogramm für die Diagnostik von Lungenrundherden. Dieses erreichte eine AUC von 0,89. Für Rundherde <2 cm zeigte das kombinierte Modell zudem eine deutlich bessere Spezifität als die alleinige Nutzung CT-Radiomics (0,90 gegen 0,67).¹⁹⁶ Kammer et al. konnten durch eine Ergänzung des unter 2.2.1.2. beschriebenen Mayo-Clinic-Modells um die Serumkonzentration eines Tumormarkers sowie einer Radiomics-Signatur eine deutliche Verbesserung der Klassifikationsleistung erreichen.¹⁹⁷ In künftigen Studien sollten klinische Biomarker berücksichtigt werden.

Die beiden Studien von Hu et al. verwendeten unterschiedliche Bildgebungsmodalitäten: native¹⁷² und kontrastmittelgestützte¹⁷³ Computertomographien. Die Studie, die native CT-Bilder verwendete, zeigte bessere Klassifikationsergebnisse.¹⁷²

Die kontrastmittelgestützte Computertomographie wurde in der vorliegenden Studie als Bildgebungsmodalität gewählt, da ein großer Teil der derzeitigen Literatur zur Klassifizierung unklarer Lungenrundherde auf Niedrigdosis-CT-Studien zu Lungenkrebscreeningzwecken basiert und die Identifikation primärer Lungenkarzinome zum Ziel hat.¹⁹⁸ In Deutschland findet ein derartiges Screening für Risikopatient*innen wie bereits erwähnt aktuell noch nicht statt¹⁹⁹, wird jedoch voraussichtlich 2026 umgesetzt.³⁰ Außerdem war das Ziel dieser Studie nicht die Diagnose von Lungenkarzinomen, sondern die Unterscheidung benigner Rundherde von Lungenmetastasen. Die kontrastmittelgestützte CT hat trotz weiter Verbreitung der Niedrigdosis-CT weiterhin diagnostischen Wert in der Untersuchung von Rundherden. Das US-amerikanische National Comprehensive Cancer Network empfiehlt in den Fällen solider Rundherde >15 mm weiterhin Folgeuntersuchungen mit Kontrastmittel.⁶² Darüber hinaus hat eine vorige Studie des radiologischen Instituts der Uniklinik Köln gezeigt, dass in der Datenbank des Instituts selbst unter ausschließlichem Einschluss von kontrastmittelgestützten DECT-Untersuchungen eine ausreichend große Kohorte möglich war.⁶⁴ Eine heterogenere Studie hatte folglich Potenzial für eine noch größere Kohorte.

Es existieren widersprüchliche Studienergebnisse bezüglich des Nutzens von Kontrastmittel für Radiomics-Analysen. Einige Studien, die kontrastmittelgestützte Untersuchungen verwendeten, zeigten signifikant schlechtere Ergebnisse im direkten Vergleich zu Untersuchungen ohne Kontrastmittel²⁰⁰ und es wurde vermutet, dass durch die Kontrastmittelanreicherung potenziell aussagekräftige Features für die Unterscheidung zwischen malignen und benignen Herden verdeckt werden könnten.²⁰¹ Gleichzeitig konnten andere Studien zeigen, dass die beiden Verfahren für die Gewinnung von Texture-Features starke Korrelationen aufweisen²⁰² und Features aus kontrastmittelgestützten Untersuchungen im Falle subsolider Rundherde sogar reproduzierbarer als die von LDCT-Untersuchungen waren.²⁰³ Außerdem zeigten Lungenrundherde-Klassifikationsstudien, die kontrastmittelgestützte CT verwendeten, überzeugende Ergebnisse (AUC = 0,941).¹⁶⁸ Ni et al. untersuchten darüber hinaus drei verschiedene Kontrastmittelphasen (nativ, arteriell und venös) und fanden die besten Klassifikationsergebnisse in der einem kombinierten Modell aller drei Phasen, gefolgt von den Ergebnissen der venösen Phase. Die native Phase zeigte sogar die schwächsten Ergebnisse der Studie.²⁰⁴ Aus diesem Grund erscheint ein signifikanter Einfluss von Kontrastmittel auf die Studienergebnisse unwahrscheinlich.

Hu et al. untersuchten zudem, ob sich signifikante Unterschiede in der Klassifikationsleistung ergeben, wenn die Radiomics-Features nur aus der der Schnittebene mit dem höchsten Durchmesser erhoben werden und nicht aus den kompletten Läsionen. Beig et al. konnten lediglich auf Basis von Features aus einer einzelnen Schnittebene vielversprechende Ergebnisse zur Klassifikation von benignen Granulomen und Adenokarzinomen erreichen (AUC 0,80).¹⁶⁶ Die AUC des Modells von Hu et al. auf Basis der Features aus den ganzen

Läsionen war jedoch höher (0,869 vs. 0,845)¹⁷³, was einen erhöhten Informationsgewinn durch vollständige Segmentierungen nahelegt. Zudem ist zu berücksichtigen, dass die Studie von Beig et al. eine vollkommen ausgeglichene Studienkohorte mit allesamt histopathologisch bestätigten Rundherden verwendete.¹⁶⁶ Wie bereits beschrieben stellen Klassenungleichgewichte und ein Mangel an definitiver histopathologischer Sicherung der Rundherdgenese Schwächen von Radiomics-Studien, unter anderem der in dieser Dissertation diskutierten Studie, dar.^{139,195} Dass die Studie trotz Meidung dieser Limitationen eine leicht geringere AUC zeigte, ist möglicherweise durch die mangelnde vollständige Herdsegmentierung zu erklären. In der vorliegenden Studie wurden daher Segmentierungen der kompletten Läsionen durchgeführt.

Özgül et al. erreichten eine AUC von 0,702, eine Sensitivität von 88 % und eine Spezifität von 70 % zur Differenzierung pulmonaler Harnblasenkarzinommetastasen von nicht-metastatischen Herden. Die Studie verwendete native Computertomographien des Thorax mit einheitlichen Scan-Protokollen, manuelle Segmentierungen der vollständigen Läsionen, und entwickelte ein Klassifikationsmodell auf Basis Shape- und Texture-Features. Als Ground Truth wurden für die Metastasen mehrheitlich eine histopathologische Sicherung verwendet, in seltenen Fällen ein positiver Befund im PET-CT.¹⁷⁰

Die Mehrheit der Radiomics-Studien zur Klassifikation von Lungenrundherden dienen jedoch weiterhin der Diagnostik von Primärmalignomen der Lunge.⁶⁴ Da dies jedoch nicht die Fragestellung der vorliegenden Studie entspricht, wird nicht explizit auf diese Studien eingegangen.

5.3.2. Rekonstruktionen

Wie bereits in Abschnitt 2.3.4. dargelegt, haben Rekonstruktionsalgorithmen Auswirkungen auf die aus den Bilddaten gewonnenen Features.^{131,132} Für eine hohe Robustheit der Features eignet sich die Verwendung einheitlicher Rekonstruktionsparameter. Da diese Studie jedoch explizit das Ziel hatte, eine heterogene Kohorte zu rekrutieren, um praxisnahe Bedingungen für ein Klassifikationsmodell zu schaffen, wurden Untersuchungen miteingeschlossen, die mit insgesamt acht verschiedenen Kernels rekonstruiert wurden (siehe Tabelle 2). Die drei häufigsten Kernels wurden daher in separaten Subgruppen analysiert. Die höchsten Sensitivitäten wurde in der Subgruppe der Iterativen Modellrekonstruktion (IMR) mit dem Kernel Sharp Plus als Lungenfenster erreicht ($0,77 \pm 0,12$ für automatisierte Feature-Selektion und $0,80 \pm 0,09$ für manuelle Feature-Selektion). Diese lagen jedoch innerhalb der Standardabweichung derer der gesamten Kohorte. Ein Grund dafür könnte sein, dass auch innerhalb der Subgruppen Daten von verschiedenen CT-Geräten verwendet wurden, was eine

weitere Quelle für Heterogenität darstellt. Die Subgruppe iDose YA zeigte zum einen die höchste Homogenität bzgl. der Untersuchungsparameter (95,4 % der Untersuchungen vom gleichen Scanner), zum anderen außerdem das geringste Klassenungleichgewicht aller Rekonstruktions-Subtypen (35,6 % benigne Rundherde, 64,4 % Metastasen). Die IMR-Subgruppe beinhaltete dagegen 25,3 % benigne Rundherde und die iDose-L-Subgruppe 19,0 % benigne Rundherde. iDose L zeigte zudem mehrheitlich die schlechtesten Klassifikationsergebnisse. Die iDose-YA-Subgruppe zeigte zwar den höchsten F1-Score sämtlicher Untersuchungen ($0,69 \pm 0,03$ und $0,69 \pm 0,01$), jedoch fiel die Sensitivität niedriger als in der IMR-Subgruppe aus. Der Grund für dieses Diskrepanz bleibt unklar.

Um dem Einfluss verschiedener Rekonstruktionskernels entgegenzuwirken, können u.a. Deep-Learning-Methoden angewandt werden. Choe et al. entwickelten auf Basis eines künstlichen neuronalen Netzes ein Modell zur Konversion der Kernels. Nach Anwendung des Modells auf Patient*innen, deren Bilddaten mit unterschiedlichen Kernels rekonstruiert wurden, wurde der CCC der Radiomics-Features erfasst und zeigte eine deutliche Verbesserung der Feature-Stabilität.¹³⁴ Derartige Modelle sollten in zukünftigen Studien mit verschiedenen Rekonstruktionen berücksichtigt werden.

5.3.3. Feature-Subgruppen

Neben der Analyse der Gesamtkohorte und der verschiedenen Rekonstruktionsalgorithmen lässt sich auch der Einfluss spezifischer Features durch Entwicklung separater Klassifikationsmodelle abschätzen.

Im Rahmen der manuellen Feature-Selektion gehörten zwei der drei selektierten Features der Kategorie der Shape-Features an (*original_shape_Flatness* und *original_shape_SurfaceVolumeRatio*). Letzteres zeigte sich bereits in anderen Studien als wertvolles Charakteristikum zur Unterscheidung von Rundherden: Zhang et al. untersuchten die Unterscheidung von benignen und malignen Lungenläsionen mithilfe von Kreuzvalidierungsverfahren. In allen fünf Kreuzvalidierungsdatensätzen wurde *SurfaceVolumeRatio* als von den Selektionsalgorithmen ausgewählt. Die Klassifikation erreichte eine AUC von bis zu $0,887 \pm 0,046$.²⁰⁵ Um den Einfluss von Shape-Features auf die Klassifikation besser einschätzen zu können, wurden sowohl ein Modell entwickelt, das ausschließlich Shape-Features verwendete als auch eins, das Shape-Features gänzlich ausschloss. Hier ergaben sich jedoch lediglich geringfügige Unterschiede der Klassifikationsleistung, die alle innerhalb der Standardabweichung der Ergebnisse der Gesamtkohorte lagen. Dies ist möglicherweise auf die Größe der untersuchten Rundherde zurückzuführen: Zhang erreichten die o.g. Ergebnisse ihrer Klassifikationsstudie ausschließlich unter Verwendung von Rundherden $>1 \text{ cm}^3$.²⁰⁵ Dieses Volumen ist größer als

die mittleren Volumina der beiden Klassen, die in dieser Studie verwendet wurden (siehe Tabelle 4). Raptis et al. konnten zudem zeigen, dass Shape-Features in kleineren Rundherden, insbesondere mit Durchmessern <2 cm, weniger aussagekräftig sind, um Malignität zu beurteilen.²⁰⁶ Daraus ließ sich schlussfolgern, dass in der vorliegenden Studie Features zu Größe und Form der Rundherde nicht von höchster Wichtigkeit für die Klassifikation waren und in Zukunft auch Texture-Features wie das verwendete *original_glcml_ClusterShade* untersucht werden sollten. Texture-Features konnten bereits erfolgreich für die Klassifikation von Lungenrundherden verwendet werden. Qi et al. erreichten beispielsweise mithilfe des *grey level run length matrix*-Features *run percentage* eine AUC von 0,954 für die Klassifikation von Lungenmetastasen. Für prämetastatische Regionen betrug die AUC 0,852.¹⁵⁸

Die bereits unter 5.3.1. beschriebene Studie von Lennartz et al. setzte sich mit einer identischen Klassifikationsaufgabe auseinander und verwendete die gleichen Einschlusskriterien wie in dieser Studie.⁶⁴ Hier wurden aus kontrastmittelgestützten Dual-Energy-CT-Untersuchungen First-order-Features aus konventionellen CT-Bildern und Iodkarten extrahiert, um ein Klassifikationsmodell zu entwickeln. Die Klassifikation der Metastasen erreichte eine Balanced Accuracy von 0,87, obwohl auch hier ein deutliches Klassenungleichgewicht vorlag (72,5 % Metastasen und 27,5 % benigne Rundherde in der Kohorte).

First-order-Features konnten außerdem von Liu et al. als wertvoll für die Identifikation von Lungenmetastasen von Kolorektalkarzinomen beschrieben werden.⁸¹

In der aktuellen Studie wurden jedoch keine First-order-Features selektiert. Für diese Diskrepanz kommen verschiedene Ursachen in Betracht. Zum einen nutzten sowohl Liu et al. als auch Lennartz et al. im Gegensatz zur vorliegenden Arbeit einen einzigen CT-Scanner mit einheitlichen Untersuchungsparametern zur Bilddatenerhebung. Wie bereits dargelegt, birgt die Verwendung eines heterogenen Datensatzes, der Bilddaten mehrerer Scanner umfasst, das Risiko einer geringeren Robustheit und Reproduzierbarkeit bestimmter Features.¹²⁹ Darüber hinaus erfolgte die Feature-Extraktion in der Studie von Liu et al. mithilfe der Software A.K. (GE Healthcare, China). Obwohl sowohl PyRadiomics als auch A.K. IBSI-konforme Features extrahieren, konnten Paquier et al. nachweisen, dass selbst standardisierte Features abhängig von der eingesetzten Software variieren können.²⁰⁷

In der vorliegenden Studie wurden ausschließlich Features der Kategorien Shape, First Order, und Texture/Second Order (*grey level co-occurrence matrix*, *grey level dependence matrix*, *grey level run length matrix* und *neighbourhood grey-tone difference matrix*) verwendet. Higher-Order-Features wurden nicht berechnet. Diese Features konnten in Studien zwar gute Klassifikationsergebnisse zeigen²⁰⁸, gelten jedoch aktuell noch als weniger

reproduzierbar.^{203,209} Aus diesem Grund wurde in dieser Studie auf die Verwendung Higher-Order-Features verzichtet.

Die vorliegende Studie nahm die Feature-Extraktion ausschließlich aus dem Volumen des Rundherdes vor, während das angrenzende Lungenparenchym nicht mitsegmentiert wurde. Frühere Arbeiten haben jedoch gezeigt, dass peritumorale Features ebenfalls einen diagnostischen Mehrwert bei der Beurteilung unklarer Lungenrundherde liefern können. So berichteten Uthoff et al., dass auf Basis ausschließlich peritumoraler Merkmale eine Sensitivität von 100 % sowie eine AUC von 0,965 zur Differenzierung benigner und maligner Rundherde erzielt werden konnte.²¹⁰ Yi et al. kombinierten intra- und peritumorale Features und erreichten eine AUC von 0,941 für die Unterscheidung benigner Rundherde von Lungenkarzinomen.¹⁶⁸

Bislang liegen nach Kenntnisstand des Autors jedoch keine publizierten Daten zur Verwendung dieser Features in Klassifikationsstudien von Lungenmetastasen vor. Dies könnte in künftigen Studien berücksichtigt werden.

5.4. Feature-Selektion

Die vorliegende Studie nutzte zwei verschiedene Methoden zur Erstellung der Klassifikationsmodelle. Diese unterschieden sich in der Art und Weise der Feature-Selektion: automatisiert im Rahmen einer Rastersuche unter Verwendung verschiedener Metriken zum Ranking sowie manuell nach individueller Testung der Klassifikationsfähigkeit und Korrelation der einzelnen Features. In der Untersuchung der gesamten Kohorte zeigten sich zwischen den beiden Ansätzen insgesamt vergleichbare Klassifikationsergebnisse, in den Subgruppen der verschiedenen Rekonstruktionsfiltern konnte das manuelle Modell vereinzelt eine bessere Klassifikationsleistung erbringen.

Die Feature-Selektion stellt einen zentralen Schritt des Radiomics-Arbeitsablaufs dar.¹⁰⁰ Ungeachtet ihrer Wichtigkeit bestehen bis heute jedoch keine allgemeingültigen Empfehlungen, welche der zahlreichen verfügbaren Methoden qualitativ am hochwertigsten ist.^{211,212} Pudjihartono et al. verglichen verschiedene Methoden und kamen zum Schluss, dass keine einzelne Methode in allen Aspekten überlegen ist und daher Kombinationen empfehlenswert seien.²¹¹ Zhang et al. konnten in ihrer Review aus dem Jahre 2023 zwar drei Metriken identifizieren die wiederholt überzeugende Ergebnisse lieferten, ohne den Arbeitsablauf unnötig zu verkomplizieren, räumten jedoch ein, dass diese Methoden nicht auf jede Forschungsfrage und jedes Studiendesign anwendbar sind.²¹²

Die automatisierte Modellentwicklung, die in der vorliegenden Studie verwendet wurde, traf keine Auswahl der errechneten Features, sondern gab lediglich vor, mit welcher Metrik eine Rangliste der Features erstellt werden sollte und wie viele der höchstplatziertesten Features für die Klassifikation genutzt werden sollten. Anders als beim Modell der manuellen Feature-Selektion variieren die Features folglich je nach Trainings-Datensatz. Dies erlaubt zwar eine noch individuellere Anpassung an die jeweilige Kohorte, birgt jedoch den Nachteil, dass eine Interpretation der verwendeten Features kaum möglich ist. Gerade im klinisch-medizinischen Kontext ist eine Interpretierbarkeit von Ergebnissen jedoch von hoher Wichtigkeit.²¹³

Es bestehen verschiedene Ansätze, um die Vorgänge und Ergebnisse maschineller Lernprozesse greifbarer und nachvollziehbarer zu machen.^{214,215} Diese liegen jedoch außerhalb des Rahmens der aktuellen Studie, sollten jedoch in der Zukunft berücksichtigt werden.

Grundsätzlich sind einfache Modelle komplexeren vorzuziehen, wenn diese gleichwertige Ergebnisse liefern.²¹³ Im Falle der aktuellen Studie erforderte die manuelle Feature-Selektion einen deutlich geringeren Arbeitsaufwand und zeigte vergleichbare Klassifikationsergebnisse. Außerdem erlaubt dieser Ansatz größere Interpretationsmöglichkeiten. Ein definitiver Vergleich der beiden Modelle sollte jedoch erst im Rahmen einer externen Validierung erfolgen.

5.5. Limitationen

Diese Studie weist mehrere Limitationen auf. Retrospektive Studien sind grundsätzlich anfällig für einen Selektionsbias. Während jeder benigner Rundherd, der die Einschlusskriterien erfüllte, in die Studienkohorte aufgenommen wurde, war dies bei den Metastasen nicht der Fall. Diese waren deutlich zahlreicher vorhanden. Aufgrund der Vorgabe von maximal zehn Rundherden pro Patient*in sowie um ein noch ausgeprägteres Klassenungleichgewicht zu vermeiden, wurde eine Vielzahl an Metastasen nicht in die Kohorte aufgenommen. Die Auswahl der Metastasen erfolgte jedoch nicht randomisiert, sodass eine Verzerrung der Studienkohorte möglich ist. Des Weiteren beruhen retrospektive Studien auf bestehenden medizinischen Daten und Informationen. Da diese nicht selbst erhoben wurden, kann nicht sichergestellt werden, dass diese vollständig oder korrekt sind. Es besteht die Gefahr eines Informationsbias.

Grundsätzlich ist für Radiomics-Studien zwar kein prospektives Design nötig, jedoch sollten prospektive Validierungsstudien angestrebt werden.²¹⁶

Verschiedene Faktoren des Studiendesigns können negative Auswirkungen auf die Klassifikationsergebnisse haben.

Eins der Ziele der Studie war die Akquise einer möglichst großen, heterogenen Kohorte, um ein auf die vielfältige klinische Praxis generalisierbares Klassifikationsmodell zu entwickeln

und zu evaluieren. Dies unterscheidet diese Studie von vielen anderen Radiomics-Studien, die unter standardisierten Bedingungen ablaufen.¹⁶⁸ Untersuchungsbedingungen wie die Auswahl des CT-Geräts, die technischen Untersuchungsparameter und die Rekonstruktionsverfahren haben jedoch großen Einfluss auf die gewonnenen Features.¹²⁹⁻¹³² Deren Robustheit ist in einer solch heterogenen Kohorte vermindert, was die Leistung des Modells reduzieren kann.^{129,132} Aus diesem Grund wurden die unter 4.2.2. erwähnten Subgruppen verschiedener Rekonstruktionskernel untersucht. In der Zukunft sollte jedoch eine externe Validierung des in dieser Studie entwickelten Modells im Vergleich mit anderen Modellen, die unter einheitlicheren Studienbedingungen entwickelt wurden¹⁶⁸, erfolgen. Insbesondere in einer ähnlich heterogenen Studienkohorte, die dem klinischen Alltag entspricht, würde dies eine Evaluation der verschiedenen Bedingungen der Modellentwicklung ermöglichen.

Wie bereits unter 5.1. erklärt, weist diese Studie ein deutliches Klassenungleichgewicht auf, was zu einem Bias des Klassifikationsmodells zur überrepräsentierten Gruppe der Metastasen führen kann. Um diesem Ungleichgewicht entgegenzuwirken, wurde Random Undersampling¹³⁹ durchgeführt. Weitere Methoden und/oder eine Quantifizierung des Effekts des Random Undersampling wurden jedoch nicht vorgenommen. In der Zukunft sollte primär ein Klassengleichgewicht angestrebt werden und bei Bedarf weitere Ansätze zur Minimierung der Effekte eines Klassenungleichgewichts wie zum Beispiel SMOTE¹⁸² ausprobiert werden. Die Mehrheit der Rundherde, die in diese Studie eingeschlossen wurden, wiesen keine histopathologische Bestätigung ihrer Dignität auf. Dies ist damit zu begründen, dass eine histologische Sicherung, selbst wenn sich bildmorphologisch der starke Verdacht einer pulmonalen Metastasierung ergibt, nur dann angestrebt wird, wenn eine Materialgewinnung aus Lungenmetastasen zur Sicherung der Diagnose oder zur Evaluation weiterführender therapeutischer Optionen nötig ist. Gutartig imponierende Rundherde werden aus Risiko-Nutzen-Abwägungen heraus noch seltener biopsiert.²¹⁷ Die Einschlusskriterien, stimmen mit einer früheren Studie, die gute Klassifikationsergebnisse erreichte⁶⁴, überein und waren für die Klassifizierung mit hoher Wahrscheinlichkeit suffizient. Ein Restrisiko einer Fehlklassifikation, z.B. durch langsam wachsende Malignome, besteht jedoch. Wie bereits erwähnt, wurde auf die histopathologische Bestätigung als Pflichtkriterium verzichtet, um eine größtmögliche Kohorte gewinnen zu können. Insbesondere benigne Rundherde werden selten biopsiert²¹⁷ und ein Ausschluss dieser Fälle hätte das bereits erwähnte Klassenungleichgewicht nur noch weiter verstärkt.

Verschiedene Aspekte der Studie reduzieren potenziell die Generalisierbarkeit des entwickelten Klassifikationsmodells. Dieses basiert auf den Daten einer einzigen Institution und schließt ausschließlich kontrastmittelgestützte CT-Untersuchungen mit ein. Darüber hinaus hat derzeit noch keine externe Validierung des Modells stattgefunden.

Verschiedene Verbesserungsansätze sollten in zukünftigen Studien berücksichtigt werden.

Das Klassifikationsmodell dieser Studie wurde ausschließlich mit intratumoralen Radiomics-Features konstruiert. Wie bereits erwähnt konnten peritumorale Features, Higher-order-Features sowie klinische Biomarker wie Tumormarker, Antikörper, Patient*innendaten und radiomorphologische Charakteristika die Klassifikationsergebnisse vieler Studien verbessern. Darüber hinaus sollten in der Zukunft subsolide und solide Rundherde sowohl gemeinsam als auch getrennt analysiert werden und die Entwicklungen bezüglich vollautomatisierter Segmentierungstechniken verfolgt werden, um den Arbeitsaufwand der Studien zu reduzieren und den Einfluss von Intra-und Interreader-Variabilität zu minimieren.

Um Radiomics-Studien zu vereinheitlichen und deren Evaluation zu erleichtern, wurde 2017 der Radiomics Quality Score entwickelt. Dieser enthält 16 Punkte zum Studiendesign und reicht von -8 bis 36 Punkten.¹²⁸ Eine Metanalyse aus dem Jahr 2025 hat gezeigt, dass seit der Einführung des Scores ein stetiger Anstieg der durchschnittlichen Punktzahl zu erkennen ist, der Score jedoch weiterhin umstritten und fehleranfällig ist. Insbesondere die Tatsache, dass externe Studienvalidierung und ein prospektives Studiendesign, welche für eine Vielzahl der durchgeführten Studien aus Kostengründen nicht möglich sind, ca. ein Drittel der Maximalpunktzahl ausmachen, wird kritisiert.²¹⁸ Diese Faktoren waren auch in dieser Studie maßgeblich für Punktzahlverlust verantwortlich. Der RQS der vorliegenden Studie betrug 11 Punkte bzw. 30,56 % (siehe Anhang 7.3.). Barry et al. berichten in Ihrer Metaanalyse einen mittleren RQS von $9,4 \pm 6,4$ bzw. $26,1 \% \pm 17,8 \%$ für sämtliche Radiomics-Studien und für Lungen-Studien einen mittleren RQS von $9,1 \pm 6,5$.²¹⁸ Kovac et al. beschreiben einen mittleren RQS für Thoraxradiologie-Studien von 36 % mit einem Konfidenzintervall von 19 % bis 42 %.²¹⁹ Insgesamt weist die vorliegende Studie eine durchschnittliche methodische Qualität auf. Eine weitere Möglichkeit zur Einschätzung der methodischen Qualität einer Radiomics-Studie stellt die CLEAR-Checkliste dar.²²⁰ Diese könnte in der Zukunft zusätzlich zur weiteren Evaluation verwendet werden.

5.6. Schlussfolgerungen

Zusammenfassend zeigt die vorliegende Arbeit, dass die Klassifikation unklarer Lungenrundherde bei onkologischen Patient*innen in benigne oder metastatisch mithilfe eines Radiomics-basierten Klassifikationsmodells in einem großen, heterogenen Datensatz mit verschiedenen CT-Geräten unterschiedlicher Hersteller und verschiedenen onkologischen Grunderkrankungen grundsätzlich möglich ist. Zum jetzigen Zeitpunkt ist die Klassifikationsleistung jedoch für eine klinische Anwendung noch unzureichend. Die explizit gewünschte Heterogenität der Studienpopulation und der Bilddaten zum Zweck einer breiten Anwendbarkeit des entwickelten Modells ging mit einer vermehrten Variabilität der Radiomics-Features einher, was die Gesamtklassifikationsleistung minderte. Der Vergleich einer

vollautomatisierten und einer manuellen Feature-Selektion im Rahmen der Modellentwicklung ergab gleichwertige bis minimal bessere Ergebnisse für die manuelle Selektion.

Zukünftige Studien könnten unter Verwendung einer ausgeglichenen Kohorte und z.B. durch Einbeziehung klinischer und laborchemischer Biomarker und peritumoraler Features die Leistung des Modells verbessern. Zudem sollten automatisierte Segmentierungstechniken zur Reduktion der Intra-/Interreader-Variabilität erwogen und eine externe Validierung des Modells angestrebt werden.

6. Literaturverzeichnis

1. Bankier AA, MacMahon H, Colby T, et al. Fleischner Society: glossary of terms for thoracic imaging. *Radiology* 2024; **310**(2): e232558.
2. van Beek EJ, Hoffman EA. Functional imaging: CT and MRI. *Clin Chest Med* 2008; **29**(1): 195-216, vii.
3. Kubo T, Ohno Y, Kauczor HU, Hatabu H. Radiation dose reduction in chest CT--review of available options. *Eur J Radiol* 2014; **83**(10): 1953-61.
4. Laurent F, Montaudon M, Corneloup O. CT and MRI of Lung Cancer. *Respiration* 2006; **73**(2): 133-42.
5. Wormanns D, Hamer OW. [Glossary of Terms for Thoracic Imaging--German Version of the Fleischner Society Recommendations]. *Rofo* 2015; **187**(8): 638-61.
6. Loverdos K, Fotiadis A, Kontogianni C, Iliopoulou M, Gaga M. Lung nodules: A comprehensive review on current approach and management. *Ann Thorac Med* 2019; **14**(4): 226-38.
7. Cruickshank A, Stieler G, Ameer F. Evaluation of the solitary pulmonary nodule. *Intern Med J* 2019; **49**(3): 306-15.
8. Choromańska A, Macura KJ. Evaluation of solitary pulmonary nodule detected during computed tomography examination. *Pol J Radiol* 2012; **77**(2): 22-34.
9. Chen D, Yang L, Zhang W, et al. Prevalence and management of pulmonary nodules: a systematic review and meta-analysis. *Journal of thoracic disease* 2024; **16**(7): 4619.
10. Das A, Bonney A, Manser R. Prevalence of pulmonary nodules detected incidentally on noncancer-related imaging: a review. *Internal Medicine Journal* 2024; **54**(9): 1440-9.
11. Li F, Sone S, Abe H, Macmahon H, Doi K. Malignant versus benign nodules at CT screening for lung cancer: comparison of thin-section CT findings. *Radiology* 2004; **233**(3): 793-8.
12. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med* 2013; **369**(10): 910-9.
13. Henschke CI, Yankelevitz DF, Mirtcheva R, McGuinness G, McCauley D, Miettinen OS. CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules. *AJR Am J Roentgenol* 2002; **178**(5): 1053-7.
14. Oda S, Awai K, Murao K, et al. Volume-doubling time of pulmonary nodules with ground glass opacity at multidetector CT: Assessment with computer-aided three-dimensional volumetry. *Acad Radiol* 2011; **18**(1): 63-9.
15. Heuvelmans MA, Walter JE, Peters RB, et al. Relationship between nodule count and lung cancer probability in baseline CT lung cancer screening: The NELSON study. *Lung Cancer* 2017; **113**: 45-50.
16. Gould MK, Tang T, Liu IL, et al. Recent Trends in the Identification of Incidental Pulmonary Nodules. *Am J Respir Crit Care Med* 2015; **192**(10): 1208-14.
17. Singhi AD, Koay EJ, Chari ST, Maitra A. Early Detection of Pancreatic Cancer: Opportunities and Challenges. *Gastroenterology* 2019; **156**(7): 2024-40.
18. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021; **71**(3): 209-49.
19. Robert Koch-Institut (Hrsg.) und die Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (Hrsg.). Krebs in Deutschland für 2019/2020. https://edoc.rki.de/bitstream/handle/176904/11438/krebs_in_deutschland_2023-1.pdf?sequence=1 (Zuletzt abgerufen am 29.03.2026)
20. Lee E, Kazerooni EA. Lung Cancer Screening. *Semin Respir Crit Care Med* 2022; **43**(6): 839-50.
21. Tockman MS. Survival and mortality from lung cancer in a screened population: the Johns Hopkins study. *Chest* 1986; **89**(4): 324S-5S.
22. Fontana RS, Sanderson DR, Woolner LB, Taylor WF, Miller WE, Muhm JR. Lung cancer screening: the Mayo program. *J Occup Med* 1986; **28**(8): 746-50.

23. Oken MM, Hocking WG, Kvale PA, et al. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. *Jama* 2011; **306**(17): 1865-73.
24. Henschke CI, McCauley DI, Yankelevitz DF, et al. Early Lung Cancer Action Project: overall design and findings from baseline screening. *Lancet* 1999; **354**(9173): 99-105.
25. Gierada DS, Black WC, Chiles C, Pinsky PF, Yankelevitz DF. Low-dose CT screening for lung cancer: evidence from 2 decades of study. *Radiology: Imaging Cancer* 2020; **2**(2): e190058.
26. Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011; **365**(5): 395-409.
27. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med* 2020; **382**(6): 503-13.
28. Becker N, Motsch E, Trotter A, et al. Lung cancer mortality reduction by LDCT screening-Results from the randomized German LUSI trial. *Int J Cancer* 2020; **146**(6): 1503-13.
29. Barkhausen J. Sie ist endlich da–die Verordnung des BMUV zur Lungenkrebsfrüherkennung. RÖFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren; 2024: Georg Thieme Verlag KG; 2024. p. 902-4.
30. Gemeinsamer Bundesausschuss. Lungenkrebs-Früherkennung bei starken Raucherinnen und Rauchern wird voraussichtlich ab April 2026 Kassenleistung. Berlin <https://www.g-ba.de/presse/pressemitteilungen-meldungen/1263/> (Zuletzt abgerufen am 29.03.2026)
31. Godoy MCB, Odisio E, Truong MT, de Groot PM, Shroff GS, Erasmus JJ. Pulmonary Nodule Management in Lung Cancer Screening: A Pictorial Review of Lung-RADS Version 1.0. *Radiol Clin North Am* 2018; **56**(3): 353-63.
32. Horeweg N, van Rosmalen J, Heuvelmans MA, et al. Lung cancer probability in patients with CT-detected pulmonary nodules: a prespecified analysis of data from the NELSON trial of low-dose CT screening. *Lancet Oncol* 2014; **15**(12): 1332-41.
33. Takashima S, Sone S, Li F, et al. Small solitary pulmonary nodules (< or =1 cm) detected at population-based CT screening for lung cancer: Reliable high-resolution CT features of benign lesions. *AJR Am J Roentgenol* 2003; **180**(4): 955-64.
34. Horeweg N, van der Aalst CM, Thunnissen E, et al. Characteristics of lung cancers detected by computer tomography screening in the randomized NELSON trial. *Am J Respir Crit Care Med* 2013; **187**(8): 848-54.
35. Mets OM, Chung K, Scholten ET, et al. Incidental perifissural nodules on routine chest computed tomography: lung cancer or not? *Eur Radiol* 2018; **28**(3): 1095-101.
36. Schreuder A, Jacobs C, Scholten ET, van Ginneken B, Schaefer-Prokop CM, Prokop M. Typical CT Features of Intrapulmonary Lymph Nodes: A Review. *Radiol Cardiothorac Imaging* 2020; **2**(4): e190159.
37. Snoeckx A, Reyntiens P, Desbuquoit D, et al. Evaluation of the solitary pulmonary nodule: size matters, but do not ignore the power of morphology. *Insights Imaging* 2018; **9**(1): 73-86.
38. Khan AN, Al-Jahdali HH, Allen CM, Irion KL, Al Ghanem S, Koteyar SS. The calcified lung nodule: What does it mean? *Ann Thorac Med* 2010; **5**(2): 67-79.
39. Webb WR. Radiologic evaluation of the solitary pulmonary nodule. *AJR Am J Roentgenol* 1990; **154**(4): 701-8.
40. Grewal RG, Austin JH. CT demonstration of calcification in carcinoma of the lung. *J Comput Assist Tomogr* 1994; **18**(6): 867-71.
41. Maile CW, Rodan BA, Godwin JD, Chen JT, Ravin CE. Calcification in pulmonary metastases. *Br J Radiol* 1982; **55**(650): 108-13.
42. Truong MT, Sabloff BS, Ko JP. Multidetector CT of solitary pulmonary nodules. *Radiol Clin North Am* 2010; **48**(1): 141-55.
43. Kim SJ, Lee KS, Ryu YH, et al. Reversed halo sign on high-resolution CT of cryptogenic organizing pneumonia: diagnostic implications. *AJR Am J Roentgenol* 2003; **180**(5): 1251-4.
44. Gould MK, Donington J, Lynch WR, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American

College of Chest Physicians evidence-based clinical practice guidelines. *Chest* 2013; **143**(5 Suppl): e93S-e120S.

45. Truong MT, Ko JP, Rossi SE, et al. Update in the evaluation of the solitary pulmonary nodule. *Radiographics* 2014; **34**(6): 1658-79.
46. Woodring JH, Fried AM, Chuang VP. Solitary cavities of the lung: diagnostic implications of cavity wall thickness. *AJR Am J Roentgenol* 1980; **135**(6): 1269-71.
47. De Cicco C, Bellomi M, Bartolomei M, et al. Imaging of lung hamartomas by multidetector computed tomography and positron emission tomography. *Ann Thorac Surg* 2008; **86**(6): 1769-72.
48. Groheux D, Quere G, Blanc E, et al. FDG PET-CT for solitary pulmonary nodule and lung cancer: Literature review. *Diagn Interv Imaging* 2016; **97**(10): 1003-17.
49. Larici AR, Farchione A, Franchi P, et al. Lung nodules: size still matters. *Eur Respir Rev* 2017; **26**(146).
50. Kobayashi Y, Fukui T, Ito S, et al. How long should small lung lesions of ground-glass opacity be followed? *J Thorac Oncol* 2013; **8**(3): 309-14.
51. Li J, Xia T, Yang X, et al. Malignant solitary pulmonary nodules: assessment of mass growth rate and doubling time at follow-up CT. *J Thorac Dis* 2018; **10**(Suppl 7): S797-s806.
52. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med* 1997; **157**(8): 849-55.
53. Gould MK, Ananth L, Barnett PG. A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. *Chest* 2007; **131**(2): 383-8.
54. Li Y, Chen KZ, Wang J. Development and validation of a clinical prediction model to estimate the probability of malignancy in solitary pulmonary nodules in Chinese people. *Clin Lung Cancer* 2011; **12**(5): 313-9.
55. Wu Z, Wang F, Cao W, et al. Lung cancer risk prediction models based on pulmonary nodules: A systematic review. *Thorac Cancer* 2022; **13**(5): 664-77.
56. Venkadesh KV, Setio AAA, Schreuder A, et al. Deep Learning for Malignancy Risk Estimation of Pulmonary Nodules Detected at Low-Dose Screening CT. *Radiology* 2021; **300**(2): 438-47.
57. Heuvelmans MA, van Ooijen PMA, Ather S, et al. Lung cancer prediction by Deep Learning to identify benign lung nodules. *Lung Cancer* 2021; **154**: 1-4.
58. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology* 2017; **284**(1): 228-43.
59. Callister ME, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax* 2015; **70** Suppl 2: ii1-ii54.
60. American College of Radiology. Lung-RADS® v2022. <https://edge.sitecorecloud.io/americancoldf5f-acrorgf92a-productioncb02-3650/media/ACR/Files/RADS/Lung-RADS/Lung-RADS-2022.pdf> (Zuletzt abgerufen am 29.03.2026)
61. Martin MD, Kanne JP, Broderick LS, Kazerooni EA, Meyer CA. Lung-RADS: Pushing the Limits. *Radiographics* 2017; **37**(7): 1975-93.
62. Wood DE, Kazerooni EA, Aberle D, et al. NCCN Guidelines® Insights: Lung Cancer Screening, Version 1.2022. *J Natl Compr Canc Netw* 2022; **20**(7): 754-64.
63. Caparica R, Mak MP, Rocha CH, et al. Pulmonary nodules in patients with nonpulmonary cancer: not always metastases. *Journal of global oncology* 2016; **2**(3): 138-44.
64. Lennartz S, Mager A, Große Hokamp N, et al. Texture analysis of iodine maps and conventional images for k-nearest neighbor classification of benign and metastatic lung nodules. *Cancer Imaging* 2021; **21**(1): 17.
65. Chen H, Stoltzfus KC, Lehrer EJ, et al. The epidemiology of lung metastases. *Frontiers in Medicine* 2021; **8**: 723396.
66. Ost D, Fein AM, Feinsilver SH. Clinical practice. The solitary pulmonary nodule. *N Engl J Med* 2003; **348**(25): 2535-42.
67. Seo JB, Im J-G, Goo JM, Chung MJ, Kim M-Y. Atypical pulmonary metastases: spectrum of radiologic findings. *Radiographics* 2001; **21**(2): 403-17.

68. Shahzad H, Ur-Rehman S, Fatima K, Sharif N, Zubairi ABS. Case series and literature review of multiple nodular sarcoidosis. *BMC Research Notes* 2013; **6**(1): 394.
69. Crow J, Slavin G, Kreel L. Pulmonary metastasis: a pathologic and radiologic study. *Cancer* 1981; **47**(11): 2595-602.
70. Gill RR, Shin Matsusoka M, Hiroto Hatabu M. Cavities in the lung in oncology patients: imaging overview and differential diagnoses. *Applied Radiology* 2010; **39**(6): 10.
71. Band PR, Kocandrl C. Growth rate of pulmonary metastases in human sarcomas. *Cancer* 1975; **36**(2): 471-4.
72. Fujimoto N, Sugita A, Terasawa Y, Kato M. Observations on the growth rate of renal cell carcinoma. *International journal of urology* 1995; **2**(2): 71-6.
73. Friberg S, Mattson S. On the growth rates of human malignant tumors: implications for medical decision making. *Journal of surgical oncology* 1997; **65**(4): 284-97.
74. Heerink WJ, de Bock GH, de Jonge GJ, Groen HJ, Vliegenthart R, Oudkerk M. Complication rates of CT-guided transthoracic lung biopsy: meta-analysis. *Eur Radiol* 2017; **27**(1): 138-48.
75. Han Y, Kim HJ, Kong KA, et al. Diagnosis of small pulmonary lesions by transbronchial lung biopsy with radial endobronchial ultrasound and virtual bronchoscopic navigation versus CT-guided transthoracic needle biopsy: A systematic review and meta-analysis. *PLoS One* 2018; **13**(1): e0191590.
76. Hwang HS, Chung MJ, Lee JW, Shin SW, Lee KS. C-arm cone-beam CT-guided percutaneous transthoracic lung biopsy: usefulness in evaluation of small pulmonary nodules. *AJR Am J Roentgenol* 2010; **195**(6): W400-7.
77. Zhang HF, Zeng XT, Xing F, Fan N, Liao MY. The diagnostic accuracy of CT-guided percutaneous core needle biopsy and fine needle aspiration in pulmonary lesions: a meta-analysis. *Clin Radiol* 2016; **71**(1): e1-10.
78. Asahina H, Yamazaki K, Onodera Y, et al. Transbronchial biopsy using endobronchial ultrasonography with a guide sheath and virtual bronchoscopic navigation. *Chest* 2005; **128**(3): 1761-5.
79. Wiener RS, Schwartz LM, Woloshin S, Welch HG. Population-based risk for complications after transthoracic needle lung biopsy of a pulmonary nodule: an analysis of discharge records. *Ann Intern Med* 2011; **155**(3): 137-44.
80. Chen CH, Chang CK, Tu CY, et al. Radiomic features analysis in computed tomography images of lung nodule classification. *PLoS One* 2018; **13**(2): e0192002.
81. Liu C, Meng Q, Zeng Q, et al. An Exploratory Study on the Stable Radiomics Features of Metastatic Small Pulmonary Nodules in Colorectal Cancer Patients. *Front Oncol* 2021; **11**: 661763.
82. Suh YJ, Lee JH, Hur J, et al. Predictors of False-Negative Results from Percutaneous Transthoracic Fine-Needle Aspiration Biopsy: An Observational Study from a Retrospective Cohort. *Yonsei Med J* 2016; **57**(5): 1243-51.
83. Philip B, Jain A, Wojtowicz M, et al. Current investigative modalities for detecting and staging lung cancers: a comprehensive summary. *Indian journal of thoracic and cardiovascular surgery* 2023; **39**(1): 42-52.
84. Zhao H, Xu Y, Huo J, Burks AC, Ost DE, Shih Y-CT. Updated analysis of complication rates associated with invasive diagnostic procedures after lung cancer screening. *JAMA Network Open* 2020; **3**(12): e2029874.
85. Mettler FA, Jr., Bhargavan M, Faulkner K, et al. Radiologic and nuclear medicine studies in the United States and worldwide: frequency, radiation dose, and comparison with other radiation sources--1950-2007. *Radiology* 2009; **253**(2): 520-31.
86. Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. *Jama* 2012; **307**(22): 2400-9.
87. Ozasa K, Shimizu Y, Suyama A, et al. Studies of the mortality of atomic bomb survivors, Report 14, 1950-2003: an overview of cancer and noncancer diseases. *Radiat Res* 2012; **177**(3): 229-43.

88. Schultz CH, Fairley R, Murphy LS, Doss M. The Risk of Cancer from CT Scans and Other Sources of Low-Dose Radiation: A Critical Appraisal of Methodologic Quality. *Prehosp Disaster Med* 2020; **35**(1): 3-16.
89. Strahlenschutzkommission I. Die Empfehlungen der Internationalen Strahlenschutzkommission (ICRP) von 2007. ICRP-Veröffentlichung 103 Verabschiedet im März 2007. *Deutsche Ausgabe März 2007*.
90. Wigge P, Schütz J. Das neue Strahlenschutzgesetz–Überblick über die Auswirkungen auf das Fachgebiet der Radiologie. *Fortschr Röntgenstr* 2017; **189**(9): 903-6.
91. Brenner DJ, Hall EJ. Computed tomography--an increasing source of radiation exposure. *N Engl J Med* 2007; **357**(22): 2277-84.
92. Mettler FA, Jr., Huda W, Yoshizumi TT, Mahesh M. Effective doses in radiology and diagnostic nuclear medicine: a catalog. *Radiology* 2008; **248**(1): 254-63.
93. Nanavaty P, Alvarez MS, Alberts WM. Lung cancer screening: advantages, controversies, and applications. *Cancer Control* 2014; **21**(1): 9-14.
94. Kerpel A, Marom EM, Green M, et al. Ultra-Low Dose Chest CT with Denoising for Lung Nodule Detection. *Isr Med Assoc J* 2021; **23**(9): 550-5.
95. Blum TG, Vogel-Claussen J, Andreas S, et al. Positionspapier zur Implementierung eines nationalen organisierten Programms in Deutschland zur Früherkennung von Lungenkrebs in Risikopopulationen mittels Low-dose-CT-Screening inklusive Management von abklärungsbedürftigen Screeningbefunden. *Pneumologie* 2024; **78**(01): 15-34.
96. Xiao R, Huang Y, Meng S, et al. A cross-sectional study of psychological burden in Chinese patients with pulmonary nodules: Prevalence and impact on the management of nodules. *Thorac Cancer* 2021; **12**(23): 3150-6.
97. Wiener RS, Gould MK, Woloshin S, Schwartz LM, Clark JA. What do you mean, a spot?: A qualitative analysis of patients' reactions to discussions with their physicians about pulmonary nodules. *Chest* 2013; **143**(3): 672-7.
98. Slatore CG, Golden SE, Ganzini L, Wiener RS, Au DH. Distress and patient-centered communication among veterans with incidental (not screen-detected) pulmonary nodules. A cohort study. *Ann Am Thorac Soc* 2015; **12**(2): 184-92.
99. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph* 2007; **31**(4-5): 198-211.
100. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging* 2020; **11**(1): 91.
101. Cheng CY, Soh ZD, Majithia S, et al. Big Data in Ophthalmology. *Asia Pac J Ophthalmol (Phila)* 2020; **9**(4): 291-8.
102. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 2012; **48**(4): 441-6.
103. Avery E, Sanelli PC, Aboian M, Payabvash S. Radiomics: A Primer on Processing Workflow and Analysis. *Semin Ultrasound CT MR* 2022; **43**(2): 142-6.
104. Lohmann P, Bousabarah K, Hoevels M, Treuer H. Radiomics in radiation oncology-basics, methods, and limitations. *Strahlenther Onkol* 2020; **196**(10): 848-55.
105. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016; **278**(2): 563-77.
106. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics* 1973; (6): 610-21.
107. Yip SS, Aerts HJ. Applications and limitations of radiomics. *Phys Med Biol* 2016; **61**(13): R150-66.
108. Choi ER, Lee HY, Jeong JY, et al. Quantitative image variables reflect the intratumoral pathologic heterogeneity of lung adenocarcinoma. *Oncotarget* 2016; **7**(41): 67302-13.
109. Moon SH, Kim J, Joung JG, et al. Correlations between metabolic texture features, genetic heterogeneity, and mutation burden in patients with lung cancer. *Eur J Nucl Med Mol Imaging* 2019; **46**(2): 446-54.
110. Sala E, Mema E, Himoto Y, et al. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clin Radiol* 2017; **72**(1): 3-10.

111. Mayerhoefer ME, Materka A, Langs G, et al. Introduction to Radiomics. *J Nucl Med* 2020; **61**(4): 488-95.
112. Wu Y, Xu L, Yang P, et al. Survival Prediction in High-grade Osteosarcoma Using Radiomics of Diagnostic Computed Tomography. *EBioMedicine* 2018; **34**: 27-34.
113. Rogers W, Thulasi Seetha S, Refaee TAG, et al. Radiomics: from qualitative to quantitative imaging. *Br J Radiol* 2020; **93**(1108): 20190948.
114. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* 2020; **295**(2): 328-38.
115. Moor J. The Dartmouth College artificial intelligence conference: The next fifty years. *Ai Magazine* 2006; **27**(4): 87-.
116. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. *Transl Vis Sci Technol* 2020; **9**(2): 14.
117. Schwendicke F, Samek W, Krois J. Artificial Intelligence in Dentistry: Chances and Challenges. *J Dent Res* 2020; **99**(7): 769-74.
118. Li Z, Kamnitsas K, Glocker B. Analyzing Overfitting Under Class Imbalance in Neural Networks for Image Segmentation. *IEEE Trans Med Imaging* 2021; **40**(3): 1065-77.
119. McBee MP, Awan OA, Colucci AT, et al. Deep Learning in Radiology. *Acad Radiol* 2018; **25**(11): 1472-80.
120. Yu G, Sun K, Xu C, et al. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. *Nat Commun* 2021; **12**(1): 6311.
121. Dayan P, Niv Y. Reinforcement learning: the good, the bad and the ugly. *Curr Opin Neurobiol* 2008; **18**(2): 185-96.
122. Ray S. A quick review of machine learning algorithms. 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon); 2019: IEEE; 2019. p. 35-9.
123. Kriegeskorte N, Golan T. Neural network models and deep learning. *Curr Biol* 2019; **29**(7): R231-r6.
124. Franceschi L, Donini M, Perrone V, et al. Hyperparameter optimization in machine learning. *arXiv preprint arXiv:241022854* 2024.
125. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* 2017; **77**(21): e104-e7.
126. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? *Science translational medicine* 2016; **8**(341): 341ps12-ps12.
127. Parvande S, Yeh H-W, Paulus MP, McKinney BA. Consensus features nested cross-validation. *Bioinformatics* 2020; **36**(10): 3093-8.
128. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017; **14**(12): 749-62.
129. Lennartz S, O'Shea A, Parakh A, Persigehl T, Baessler B, Kambadakone A. Robustness of dual-energy CT-derived radiomic features across three different scanner types. *Eur Radiol* 2022; **32**(3): 1959-70.
130. Park S, Lee SM, Do KH, et al. Deep Learning Algorithm for Reducing CT Slice Thickness: Effect on Reproducibility of Radiomic Features in Lung Cancer. *Korean J Radiol* 2019; **20**(10): 1431-40.
131. Rizzo S, Botta F, Raimondi S, et al. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp* 2018; **2**(1): 36.
132. Mackin D, Ger R, Gay S, et al. Matching and Homogenizing Convolution Kernels for Quantitative Studies in Computed Tomography. *Invest Radiol* 2019; **54**(5): 288-95.
133. Shafiq-UI-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys* 2017; **44**(3): 1050-62.
134. Choe J, Lee SM, Do KH, et al. Deep Learning-based Image Conversion of CT Reconstruction Kernels Improves Radiomics Reproducibility for Pulmonary Nodules or Masses. *Radiology* 2019; **292**(2): 365-73.
135. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. *Radiology* 2019; **291**(1): 53-9.

136. Mali SA, Ibrahim A, Woodruff HC, et al. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. *J Pers Med* 2021; **11**(9).
137. Kocak B, Durmaz ES, Kaya OK, Ates E, Kilickesmez O. Reliability of Single-Slice-Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility. *AJR Am J Roentgenol* 2019; **213**(2): 377-83.
138. Barragán-Montero A, Bibal A, Dastarac MH, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. *Phys Med Biol* 2022; **67**(11).
139. Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. 2020 11th international conference on information and communication systems (ICICS); 2020: IEEE; 2020. p. 243-8.
140. Chen M, Linton-Reid K, Aboagye EO, Copley SJ. Translating radiomics into clinical practice: A step-by-step guide to study design and evaluation. *Clinical Radiology* 2025; **90**: 107053.
141. Afarid M, Mohsenipoor N, Parsaei H, et al. Assessment of macular findings by OCT angiography in patients without clinical signs of diabetic retinopathy: radiomics features for early screening of diabetic retinopathy. *BMC Ophthalmol* 2022; **22**(1): 281.
142. Tafuri B, Filardi M, Urso D, et al. Radiomics Model for Frontotemporal Dementia Diagnosis Using T1-Weighted MRI. *Front Neurosci* 2022; **16**: 828029.
143. Luo R, Gao J, Gan W, Xie WB. Clinical-radiomics nomogram for predicting esophagogastric variceal bleeding risk noninvasively in patients with cirrhosis. *World J Gastroenterol* 2023; **29**(6): 1076-89.
144. Li H, Mo Y, Huang C, et al. An MSCT-based radiomics nomogram combined with clinical factors can identify Crohn's disease and ulcerative colitis. *Ann Transl Med* 2021; **9**(7): 572.
145. Wilson R, Devaraj A. Radiomics of pulmonary nodules and lung cancer. *Transl Lung Cancer Res* 2017; **6**(1): 86-91.
146. Liu Z, Wang S, Dong D, et al. The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. *Theranostics* 2019; **9**(5): 1303-22.
147. Chu LC, Park S, Kawamoto S, et al. Utility of CT Radiomics Features in Differentiation of Pancreatic Ductal Adenocarcinoma From Normal Pancreatic Tissue. *AJR Am J Roentgenol* 2019; **213**(2): 349-57.
148. De Diego IM, Redondo AR, Fernández RR, Navarro J, Moguerza JM. General Performance Score for classification problems. *Applied Intelligence* 2022; **52**(10): 12049-63.
149. Fawcett T. An introduction to ROC analysis. *Pattern recognition letters* 2006; **27**(8): 861-74.
150. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. *Lancet Oncol* 2015; **16**(4): e173-80.
151. Zhu X, Dong D, Chen Z, et al. Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. *Eur Radiol* 2018; **28**(7): 2772-8.
152. Ba-Ssalamah A, Muin D, Scherthaner R, et al. Texture-based classification of different gastric tumors at contrast-enhanced CT. *Eur J Radiol* 2013; **82**(10): e537-43.
153. Liu S, Liu S, Ji C, et al. Application of CT texture analysis in predicting histopathological characteristics of gastric cancers. *Eur Radiol* 2017; **27**(12): 4951-9.
154. Wang G, Wang B, Wang Z, et al. Radiomics signature of brain metastasis: prediction of EGFR mutation status. *Eur Radiol* 2021; **31**(7): 4538-47.
155. Yang L, Dong D, Fang M, et al. Can CT-based radiomics signature predict KRAS/NRAS/BRAF mutations in colorectal cancer? *Eur Radiol* 2018; **28**(5): 2058-67.
156. Maldonado F, Boland JM, Raghunath S, et al. Noninvasive characterization of the histopathologic features of pulmonary nodules of the lung adenocarcinoma spectrum using computer-aided nodule assessment and risk yield (CANARY)--a pilot study. *J Thorac Oncol* 2013; **8**(4): 452-60.

157. Dong Y, Feng Q, Yang W, et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI. *Eur Radiol* 2018; **28**(2): 582-91.
158. Qi Y, Cui X, Han M, et al. Radiomics analysis of lung CT image for the early detection of metastases in patients with breast cancer: preliminary findings from a retrospective cohort study. *Eur Radiol* 2020; **30**(8): 4545-56.
159. Coroller TP, Grossmann P, Hou Y, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol* 2015; **114**(3): 345-50.
160. Horvat N, Veeraraghavan H, Khan M, et al. MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy. *Radiology* 2018; **287**(3): 833-43.
161. Mattonen SA, Palma DA, Johnson C, et al. Detection of Local Cancer Recurrence After Stereotactic Ablative Radiation Therapy for Lung Cancer: Physician Performance Versus Radiomic Assessment. *Int J Radiat Oncol Biol Phys* 2016; **94**(5): 1121-8.
162. Hawkins S, Wang H, Liu Y, et al. Predicting malignant nodules from screening CT scans. *Journal of Thoracic Oncology* 2016; **11**(12): 2120-8.
163. Warkentin MT, Al-Sawaihey H, Lam S, et al. Radiomics analysis to predict pulmonary nodule malignancy using machine learning approaches. *Thorax* 2024; **79**(4): 307-15.
164. Liu J, Xu H, Qing H, et al. Comparison of Radiomic Models Based on Low-Dose and Standard-Dose CT for Prediction of Adenocarcinomas and Benign Lesions in Solid Pulmonary Nodules. *Front Oncol* 2020; **10**: 634298.
165. Zhao J, Sun L, Sun K, et al. Development and Validation of a Radiomics Nomogram for Differentiating Pulmonary Cryptococcosis and Lung Adenocarcinoma in Solitary Pulmonary Solid Nodule. *Front Oncol* 2021; **11**: 759840.
166. Beig N, Khorrami M, Alilou M, et al. Perinodular and Intranodular Radiomic Features on Lung CT Images Distinguish Adenocarcinomas from Granulomas. *Radiology* 2019; **290**(3): 783-92.
167. Dilger SK, Uthoff J, Judisch A, et al. Improved pulmonary nodule classification utilizing quantitative lung parenchyma features. *Journal of Medical Imaging* 2015; **2**(4): 041004-.
168. Yi L, Peng Z, Chen Z, et al. Identification of pulmonary adenocarcinoma and benign lesions in isolated solid lung nodules based on a nomogram of intranodal and perinodal CT radiomic features. *Front Oncol* 2022; **12**: 924055.
169. Gabelloni M, Faggioni L, Fusco R, et al. Radiomics in Lung Metastases: A Systematic Review. *J Pers Med* 2023; **13**(2).
170. Özgül HA, Altay C, Başara Akin I, et al. [Evaluation of solid pulmonary nodules in patients with bladder cancer with computed tomography based texture analysis method: Is it possible to differentiate metastatic - non-metastatic nodules?]. *Tuberk Toraks* 2021; **69**(2): 227-36.
171. Cho YJ, Kim WS, Choi YH, et al. Computerized texture analysis of pulmonary nodules in pediatric patients with osteosarcoma: Differentiation of pulmonary metastases from non-metastatic nodules. *PLoS One* 2019; **14**(2): e0211969.
172. Hu T, Wang S, Huang L, et al. A clinical-radiomics nomogram for the preoperative prediction of lung metastasis in colorectal cancer patients with indeterminate pulmonary nodules. *Eur Radiol* 2019; **29**(1): 439-49.
173. Hu T, Wang S, E X, et al. CT Morphological Features Integrated With Whole-Lesion Histogram Parameters to Predict Lung Metastasis for Colorectal Cancer Patients With Pulmonary Nodules. *Front Oncol* 2019; **9**: 1241.
174. Ge G, Zhang J. Feature selection methods and predictive models in CT lung cancer radiomics. *J Appl Clin Med Phys* 2023; **24**(1): e13869.
175. Wand M. Data-based choice of histogram bin width. *The American Statistician* 1997; **51**(1): 59-64.
176. Lavrova E, Primakov S, Salahuddin Z, et al. Precision-medicine-toolbox: An open-source python package for the quantitative medical image analysis. *Software Impacts* 2023; **16**: 100508.
177. Domingos P. A few useful things to know about machine learning. *Communications of the ACM* 2012; **55**(10): 78-87.

178. Chicco D. Ten quick tips for machine learning in computational biology. *BioData mining* 2017; **10**(1): 35.
179. Wongvorachan T, He S, Bulut O. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information* 2023; **14**(1): 54.
180. He H, Ma Y. Imbalanced learning: foundations, algorithms, and applications. 2013.
181. Weiss GM. Foundations of imbalanced learning. *Imbalanced learning: Foundations, algorithms, and applications* 2013: 13-41.
182. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *The Journal of artificial intelligence research* 2011; **16**: 321-57.
183. Haluška R, Brabec J, Komárek T. Benchmark of data preprocessing methods for imbalanced classification. 2022 IEEE International Conference on Big Data (Big Data); 2022: IEEE; 2022. p. 2970-9.
184. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. *Radiology* 2008; **246**(3): 697-722.
185. Shimizu K, Ikeda N, Tsuboi M, Hirano T, Kato H. Percutaneous CT-guided fine needle aspiration for lung cancer smaller than 2 cm and revealed by ground-glass opacity at CT. *Lung Cancer* 2006; **51**(2): 173-9.
186. Jansen MJA, Kuijff HJ, Veldhuis WB, Wessels FJ, Viergever MA, Pluim JPW. Automatic classification of focal liver lesions based on MRI and risk factors. *PLoS One* 2019; **14**(5): e0217053.
187. Gatos I, Tsantis S, Karamesini M, et al. Focal liver lesions segmentation and classification in nonenhanced T2-weighted MRI. *Med Phys* 2017; **44**(7): 3695-705.
188. Kim YT. Management of Ground-Glass Nodules: When and How to Operate? *Cancers (Basel)* 2022; **14**(3).
189. Dalpiaz G, Asioli S, Fanti S, Rea G, Marchiori E. Rapidly growing pulmonary ground-glass nodule caused by metastatic melanoma lacking uptake on 18F-FDG PET-CT. *J Bras Pneumol* 2018; **44**(2): 171-2.
190. Nicholas E, Braff S, Klein JS. Evaluation of the solitary pulmonary nodule: A practical approach. *Applied Radiology* 2011; **40**(12): 6.
191. Wu FZ, Wu YJ, Tang EK. An integrated nomogram combined semantic-radiomic features to predict invasive pulmonary adenocarcinomas in subjects with persistent subsolid nodules. *Quant Imaging Med Surg* 2023; **13**(2): 654-68.
192. Digumarthy SR, Padole AM, Rastogi S, et al. Predicting malignant potential of subsolid nodules: can radiomics preempt longitudinal follow up CT? *Cancer Imaging* 2019; **19**(1): 36.
193. Xiong Z, Xia Q, Hu Z, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Med Image Anal* 2021; **67**: 101832.
194. Baeßler B, Weiss K, Pinto Dos Santos D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Invest Radiol* 2019; **54**(4): 221-8.
195. Hu X, Ye W, Li Z, et al. Non-invasive evaluation for benign and malignant subcentimeter pulmonary ground-glass nodules (≤ 1 cm) based on CT texture analysis. *Br J Radiol* 2020; **93**(1114): 20190762.
196. Ding Y, Zhang J, Zhuang W, et al. Improving the efficiency of identifying malignant pulmonary nodules before surgery via a combination of artificial intelligence CT image recognition and serum autoantibodies. *Eur Radiol* 2022.
197. Kammer MN, Lakhani DA, Balar AB, et al. Integrated Biomarkers for the Management of Indeterminate Pulmonary Nodules. *Am J Respir Crit Care Med* 2021; **204**(11): 1306-16.
198. Sun Y, Zhao W, Kuang K, et al. Non-contrast and contrast enhanced computed tomography radiomics in preoperative discrimination of lung invasive and non-invasive adenocarcinoma. *Front Med (Lausanne)* 2022; **9**: 939434.
199. Kaaks R, Delorme S. Screening von Bronchial-und Lungenkrebs. *Kompandium Internistische Onkologie: Standards in Diagnostik und Therapie* 2020: 1-7.
200. Dennie C, Thornhill R, Sethi-Virmani V, et al. Role of quantitative computed tomography texture analysis in the differentiation of primary lung cancer and granulomatous nodules. *Quant Imaging Med Surg* 2016; **6**(1): 6-15.

201. Dujardin M, Gibbs P, Turnbull LW. Texture analysis of 3T high resolution T2 weighted images in ovarian cystadenoma versus borderline tumor. *Proc Intl Soc Magn Reson Med*; 2014; 2014. p. 2218.
202. Ganeshan B, Goh V, Mandeville HC, Ng QS, Hoskin PJ, Miles KA. Non-small cell lung cancer: histopathologic correlates for texture parameters at CT. *Radiology* 2013; **266**(1): 326-36.
203. Gao Y, Hua M, Lv J, et al. Reproducibility of radiomic features of pulmonary nodules between low-dose CT and conventional-dose CT. *Quant Imaging Med Surg* 2022; **12**(4): 2368-77.
204. Ni XQ, Yin HK, Fan GH, Shi D, Xu L, Jin D. Differentiation of pulmonary sclerosing pneumocytoma from solid malignant pulmonary nodules by radiomic analysis on multiphasic CT. *J Appl Clin Med Phys* 2021; **22**(2): 158-64.
205. Zhang R, Zhu L, Cai Z, et al. Potential feature exploration and model development based on 18F-FDG PET/CT images for differentiating benign and malignant lung lesions. *Eur J Radiol* 2019; **121**: 108735.
206. Raptis S, Ilioudis C, Theodorou K. Uncovering the diagnostic power of radiomic feature significance in automated lung cancer detection: An integrative analysis of texture, shape, and intensity contributions. *BioMedInformatics* 2024; **4**(4): 2400-25.
207. Paquier Z, Chao S-L, Acquisto A, et al. Radiomics software comparison using digital phantom and patient data: IBSI-compliance does not guarantee concordance of feature values. *Biomedical Physics & Engineering Express* 2022; **8**(6): 065008.
208. Jing R, Wang J, Li J, et al. A wavelet features derived radiomics nomogram for prediction of malignant and benign early-stage lung nodules. *Sci Rep* 2021; **11**(1): 22330.
209. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int J Radiat Oncol Biol Phys* 2018; **102**(4): 1143-58.
210. Uthoff J, Stephens MJ, Newell JD, Jr., et al. Machine learning approach for distinguishing malignant and benign lung nodules utilizing standardized perinodular parenchymal features from CT. *Med Phys* 2019; **46**(7): 3207-16.
211. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in bioinformatics* 2022; **2**: 927312.
212. Zhang W, Guo Y, Jin Q. Radiomics and its feature selection: a review. *Symmetry* 2023; **15**(10): 1834.
213. Hong S, Hong S, Oh E, Lee WJ, Jeong WK, Kim K. Development of a flexible feature selection framework in radiomics-based prediction modeling: Assessment with four real-world datasets. *Scientific Reports* 2024; **14**(1): 29297.
214. Lopuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* 2019; **10**(1): 1096.
215. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 2017; **30**.
216. Lubner MG, Smith AD, Sandrasegaran K, Sahani DV, Pickhardt PJ. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. *Radiographics* 2017; **37**(5): 1483-503.
217. Liu A, Wang Z, Yang Y, et al. Preoperative diagnosis of malignant pulmonary nodules in lung cancer screening with a radiomics nomogram. *Cancer Commun (Lond)* 2020; **40**(1): 16-24.
218. Barry N, Kendrick J, Molin K, et al. Evaluating the impact of the Radiomics Quality Score: a systematic review and meta-analysis. *European radiology* 2025; **35**(3): 1701-13.
219. Kocak B, Keles A, Kose F, Sendur A. Quality of radiomics research: comprehensive analysis of 1574 unique publications from 89 reviews. *European Radiology* 2025; **35**(4): 1980-92.
220. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMI. *Insights into imaging* 2023; **14**(1): 75.

7. Anhang

7.1. Abbildungsverzeichnis

Abbildung 1: Selektion der Studienkohorte	40
Abbildung 2: Übersicht des Arbeitsablaufes der Klassifikationsmodellerstellung. A: Übersicht der Modellentwicklungsprozesses unter Verwendung von Kreuzvalidierung. B: Detaillierte Darstellung der automatisierten Modellentwicklung. C: Detaillierte Darstellung der manuellen Modellentwicklung.	45
Abbildung 3: Interreader-Variabilitätsanalyse. Dice-Score in Relation zum Rundherd-Volumen.	49
Abbildung 4: Interreader-Variabilitätsanalyse. Konkordanz-Korrelationskoeffizient in Relation zum Rundherd-Volumen	50
Abbildung 5: Klassifikationsergebnisse innerhalb der gesamten Studienkohorte unter Verwendung der automatisierten Feature-Selektion. ROC-Kurve der einzelnen äußeren Test-Datenätze und Darstellung der Falsch-Positiv-Rate (1-Spezifität) und Echt-Positiv-Rate (Sensitivität)	52
Abbildung 6: Klassifikationsergebnisse innerhalb der gesamten Studienkohorte unter Verwendung der manuell extrahierten Features nach explorativer Datenanalyse. ROC-Kurve der einzelnen äußeren Test-Datenätze und Darstellung der Falsch-Positiv-Rate (1-Spezifität) und Echt-Positiv-Rate (Sensitivität).	52

7.2. Tabellenverzeichnis

Tabelle 1: Differenzialdiagnosen von Lungenrundherden	13
Tabelle 2: Bildgewinnungs- und Rekonstruktionsparameter	42
Tabelle 3: Patient*innencharakteristika	48
Tabelle 4: Größe der Rundherde, berechnet durch die entsprechenden Shape-Features. Angaben in Mittelwert $\pm$ Standardabweichung.	48
Tabelle 5: Klassifikationsergebnisse des Ensemble-Klassifikators unter Verwendung des automatisierten Arbeitsablaufes zur Klassifikatorentwicklung. Angaben in Mittelwert $\pm$ Standardabweichung.	51
Tabelle 6: Klassifikationsergebnisse des Ensemble-Klassifikators unter Verwendung des manuellen Arbeitsablaufes zur Klassifikatorentwicklung. Aufgrund der manuell selektierten Features ist keine Testung der Feature-Subkategorien möglich. Angaben in Mittelwert $\pm$ Standardabweichung	51

7.3. Radiomics Quality Score

Der nachfolgende Fragebogen entspricht dem von Lambin et al. entwickelten Radiomics Quality Score (RQS)¹²⁸ und ist der vorliegenden Studie entsprechend ausgefüllt und ausgewertet.

Image protocol quality - well-documented image protocols (for example, contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability

• ☒

• ☐

• ☐

protocols well documented

public protocol used

none

Multiple segmentations - possible actions are: segmentation by different physicians/ algorithms/ software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyse feature robustness to segmentation variabilities

• ☒

• ☐

yes

no

Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyse feature robustness to these sources of variability

• ☐

• ☒

yes

no

Imaging at multiple time points - collect images of individuals at additional time points. Analyse feature robustness to temporal variabilities (for example, organ movement, organ expansion/shrinkage)

• ☐

• ☒

yes

no

Feature reduction or adjustment for multiple testing - decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features

• ☒

• ☐

Either measure is implemented

Neither measure is implemented

Multivariable analysis with non radiomics features (for example, EGFR mutation) - is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non radiomics features

• ☐

• ☒

yes

no

Detect and discuss biological correlates - demonstration of phenotypic differences (possibly associated with underlying gene–protein expression patterns) deepens understanding of radiomics and biology

• ☐

• ☒

yes

no

Cut-off analyses - determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results

• ☐

• ☒

yes

no

Discrimination statistics - report discrimination statistics (for example, C-statistic, ROC curve, AUC) and their statistical significance (for example, p-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)

• ☒

• ☒

• ☐

a discrimination statistic and its statistical significance are reported

a resampling method technique is also applied

none

Calibration statistics - report calibration statistics (for example, Calibration-in-the-large/slope, calibration plots) and their statistical significance (for example, P-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)

• ☐

• ☒

• ☐

a calibration statistic and its statistical significance are reported

a resampling method technique is applied

none

Prospective study registered in a trial database - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker

• ☐

• ☒

yes

no

Validation - the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance

• ☐

• ☒

• ☐

No validation

validation is based on a dataset from the same institute

validation is based on a dataset from another institute

• ☐

• ☐

• ☐

validation is based on two datasets from

the study validates a previously published signature

validation is based on three or more datasets from distinct institutes

two distinct
institutes

Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM-staging for survival prediction). This comparison shows the added value of radiomics

- | | |
|----------------------------|---------------------------------------|
| • ☐ | • ☒ |
| yes | no |

Potential clinical utility - report on the current and potential application of the model in a clinical setting (for example, decision curve analysis).

- | | |
|----------------------------|---------------------------------------|
| • ☐ | • ☒ |
| yes | no |

Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (for example, QALYs generated)

- | | |
|----------------------------|---------------------------------------|
| • ☐ | • ☒ |
| yes | no |

Open science and data - make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study

- | | | | |
|----------------------------|--|---------------------------------------|---|
| • ☐ | • ☐ | • ☒ | • ☐ |
| scans are open source | region of interest segmentations are open source | the code is open sourced | radiomics features are calculated on a set of representative ROIs and the calculated features and representative ROIs are open source |

Gesamtpunktzahl: 11 (30.56 %)

7.4. Konkordanz-Korrelationskoeffizienten der Radiomics-Features innerhalb der Kohorte der Intrareader-Variabilitätsanalyse

Konkordanz-Korrelationskoeffizienten (CCC) <0,75 führten zum Ausschluss der Features.

Feature	CCC	Einschluss
original_shape_Elongation	0.48 (0.32 - 0.62)	nein
original_shape_Flatness	0.75 (0.65 - 0.83)	ja
original_shape_LeastAxisLength	0.92 (0.88 - 0.95)	ja
original_shape_MajorAxisLength	0.88 (0.82 - 0.92)	ja
original_shape_Maximum2DDiameterColumn	0.86 (0.80 - 0.91)	ja
original_shape_Maximum2DDiameterRow	0.83 (0.76 - 0.88)	ja

original_shape_Maximum2DDiameterSlice	0.91 (0.86 - 0.94)	ja
original_shape_Maximum3DDiameter	0.86 (0.81 - 0.91)	ja
original_shape_MeshVolume	0.95 (0.92 - 0.96)	ja
original_shape_MinorAxisLength	0.89 (0.84 - 0.92)	ja
original_shape_Sphericity	0.61 (0.48 - 0.72)	nein
original_shape_SurfaceArea	0.92 (0.89 - 0.95)	ja
original_shape_SurfaceVolumeRatio	0.82 (0.74 - 0.88)	ja
original_shape_VoxelVolume	0.94 (0.92 - 0.96)	ja
original_firstorder_10Percentile	0.36 (0.21 - 0.50)	nein
original_firstorder_90Percentile	0.98 (0.97 - 0.98)	ja
original_firstorder_Energy	0.94 (0.92 - 0.96)	ja
original_firstorder_Entropy	0.50 (0.34 - 0.64)	nein
original_firstorder_InterquartileRange	0.70 (0.58 - 0.79)	nein
original_firstorder_Kurtosis	0.68 (0.56 - 0.77)	nein
original_firstorder_Maximum	0.97 (0.96 - 0.98)	ja
original_firstorder_MeanAbsoluteDeviation	0.78 (0.68 - 0.84)	ja
original_firstorder_Mean	0.81 (0.75 - 0.86)	ja
original_firstorder_Median	0.78 (0.71 - 0.84)	ja
original_firstorder_Minimum	0.08 (0.21 - 0.05)	nein
original_firstorder_Range	0.84 (0.77 - 0.89)	ja
original_firstorder_RobustMeanAbsoluteDeviation	0.74 (0.63 - 0.82)	nein
original_firstorder_RootMeanSquared	0.74 (0.65 - 0.81)	nein
original_firstorder_Skewness	0.71 (0.60 - 0.79)	nein
original_firstorder_TotalEnergy	0.97 (0.95 - 0.98)	ja
original_firstorder_Uniformity	0.44 (0.28 - 0.58)	nein
original_firstorder_Variance	0.95 (0.92 - 0.96)	ja
original_glcm_Autocorrelation	0.82 (0.75 - 0.87)	ja
original_glcm_JointAverage	0.69 (0.58 - 0.77)	nein
original_glcm_ClusterProminence	0.99 (0.99 - 0.99)	ja
original_glcm_ClusterShade	0.96 (0.95 - 0.97)	ja
original_glcm_ClusterTendency	0.96 (0.93 - 0.97)	ja
original_glcm_Contrast	0.95 (0.93 - 0.97)	ja
original_glcm_Correlation	0.58 (0.44 - 0.70)	nein
original_glcm_DifferenceAverage	0.85 (0.79 - 0.90)	ja
original_glcm_DifferenceEntropy	0.46 (0.29 - 0.60)	nein
original_glcm_DifferenceVariance	0.95 (0.93 - 0.97)	ja
original_glcm_JointEnergy	0.45 (0.30 - 0.58)	nein
original_glcm_JointEntropy	0.78 (0.69 - 0.85)	ja
original_glcm_Imc1	0.97 (0.96 - 0.98)	ja
original_glcm_Imc2	0.78 (0.70 - 0.83)	ja
original_glcm_Idm	0.87 (0.82 - 0.91)	ja
original_glcm_Idmn	0.83 (0.76 - 0.88)	ja
original_glcm_Id	0.84 (0.77 - 0.89)	ja
original_glcm_Idn	0.88 (0.83 - 0.91)	ja
original_glcm_InverseVariance	0.85 (0.79 - 0.90)	ja
original_glcm_MaximumProbability	0.48 (0.33 - 0.61)	nein
original_glcm_SumEntropy	0.56 (0.41 - 0.68)	nein
original_glcm_SumSquares	0.96 (0.93 - 0.97)	ja
original_gldm_DependenceEntropy	0.76 (0.66 - 0.83)	ja
original_gldm_DependenceNonUniformity	0.86 (0.80 - 0.91)	ja

original_gldm_DependenceNonUniformityNormalized	0.91 (0.87 - 0.94)	ja
original_gldm_DependenceVariance	0.99 (0.98 - 0.99)	ja
original_gldm_GrayLevelNonUniformity	0.99 (0.98 - 0.99)	ja
original_gldm_GrayLevelVariance	0.95 (0.92 - 0.96)	ja
original_gldm_HighGrayLevelEmphasis	0.84 (0.78 - 0.89)	ja
original_gldm_LargeDependenceEmphasis	0.96 (0.95 - 0.97)	ja
original_gldm_LargeDependenceHighGrayLevelEmphasis	0.79 (0.74 - 0.83)	ja
original_gldm_LargeDependenceLowGrayLevelEmphasis	0.44 (0.27 - 0.58)	nein
original_gldm_LowGrayLevelEmphasis	0.60 (0.47 - 0.71)	nein
original_gldm_SmallDependenceEmphasis	0.90 (0.85 - 0.93)	ja
original_gldm_SmallDependenceHighGrayLevelEmphasis	0.90 (0.86 - 0.93)	ja
original_gldm_SmallDependenceLowGrayLevelEmphasis	0.57 (0.43 - 0.68)	nein
original_glrlm_GrayLevelNonUniformity	0.99 (0.98 - 0.99)	ja
original_glrlm_GrayLevelNonUniformityNormalized	0.43 (0.27 - 0.57)	nein
original_glrlm_GrayLevelVariance	0.95 (0.92 - 0.96)	ja
original_glrlm_HighGrayLevelRunEmphasis	0.84 (0.78 - 0.89)	ja
original_glrlm_LongRunEmphasis	0.95 (0.93 - 0.96)	ja
original_glrlm_LongRunHighGrayLevelEmphasis	0.83 (0.76 - 0.88)	ja
original_glrlm_LongRunLowGrayLevelEmphasis	0.60 (0.46 - 0.71)	nein
original_glrlm_LowGrayLevelRunEmphasis	0.60 (0.46 - 0.71)	nein
original_glrlm_RunEntropy	0.54 (0.38 - 0.67)	nein
original_glrlm_RunLengthNonUniformity	0.94 (0.91 - 0.96)	ja
original_glrlm_RunLengthNonUniformityNormalized	0.94 (0.91 - 0.96)	ja
original_glrlm_RunPercentage	0.95 (0.93 - 0.96)	ja
original_glrlm_RunVariance	0.96 (0.95 - 0.97)	ja
original_glrlm_ShortRunEmphasis	0.94 (0.91 - 0.95)	ja
original_glrlm_ShortRunHighGrayLevelEmphasis	0.85 (0.78 - 0.89)	ja
original_glrlm_ShortRunLowGrayLevelEmphasis	0.60 (0.46 - 0.71)	nein
original_glszm_GrayLevelNonUniformity	0.98 (0.97 - 0.99)	ja
original_glszm_GrayLevelNonUniformityNormalized	0.49 (0.33 - 0.62)	nein
original_glszm_GrayLevelVariance	0.94 (0.92 - 0.96)	ja
original_glszm_HighGrayLevelZoneEmphasis	0.86 (0.80 - 0.90)	ja
original_glszm_LargeAreaEmphasis	0.87 (0.87 - 0.88)	ja
original_glszm_LargeAreaHighGrayLevelEmphasis	0.95 (0.94 - 0.95)	ja
original_glszm_LargeAreaLowGrayLevelEmphasis	0.50 (0.34 - 0.63)	nein
original_glszm_LowGrayLevelZoneEmphasis	0.55 (0.41 - 0.67)	nein
original_glszm_SizeZoneNonUniformity	0.85 (0.79 - 0.90)	ja
original_glszm_SizeZoneNonUniformityNormalized	0.84 (0.77 - 0.89)	ja
original_glszm_SmallAreaEmphasis	0.83 (0.76 - 0.89)	ja
original_glszm_SmallAreaHighGrayLevelEmphasis	0.88 (0.83 - 0.92)	ja
original_glszm_SmallAreaLowGrayLevelEmphasis	0.52 (0.37 - 0.65)	nein
original_glszm_ZoneEntropy	0.70 (0.59 - 0.79)	nein
original_glszm_ZonePercentage	0.91 (0.87 - 0.94)	ja
original_glszm_ZoneVariance	0.87 (0.87 - 0.88)	ja
original_ngtdm_Busyness	0.36 (0.24 - 0.46)	nein
original_ngtdm_Coarseness	0.61 (0.47 - 0.72)	nein
original_ngtdm_Complexity	0.93 (0.92 - 0.95)	ja
original_ngtdm_Contrast	0.80 (0.73 - 0.86)	ja
original_ngtdm_Strength	0.89 (0.85 - 0.92)	ja

