Aus der Klinik und Poliklinik für Radiologie
Klinikum der Ludwig-Maximilians-Universität München

Direktor: Prof. Dr. med. Jens Ricke

Klinische Validierung, Optimierung und Wahrnehmung von Algorithmen der künstlichen Intelligenz in der diagnostischen Radiologie

Kumulative Habilitationsschrift
zur Erlangung der Venia Legendi
im Fach „Radiologie“

vorgelegt von
Dr. med. Jan Philipp Rudolph
aus Göttingen

2026

Inhaltsverzeichnis

1 Einleitung

Künstliche Intelligenz (KI) mit ihren potentiell vielseitigen Anwendungsmöglichkeiten in der diagnostischen Bildgebung hat in den vergangenen Jahren ein enormes Interesse sowohl bei Informatikern/Data Scientists, als auch bei Radiologen, klinisch tätigen Medizinern und Patienten geweckt.[1] Nach allgemeiner Auffassung wurde der heute übliche englische Begriff für künstliche Intelligenz „artificial intelligence“ zum ersten Mal im Rahmen der Dartmouth-College-Konferenz im Jahr 1956 verwendet [1]. Es dauerte eine Dekade bis erste Anwendungsmöglichkeiten in den Biowissenschaften auftauchten, vor allem prominent im Rahmen der DENDRAL-Experimente der späten 1960er und frühen 1970er Jahre, bei denen sich Informatiker, Chemiker und Philosophen kollaborativ zusammenschlossen und Computerprogramme mit speziellem (insbesondere chemischem) Expertenwissen fütterten. Dabei wurden die Programme meist mit einem heuristischen Ansatz genutzt, um generalisierte Schlüsse – wie z. B. die Formulierung einer wissenschaftlichen Hypothese – zu ziehen [2, 3]. Die Entwicklung erster radiologischer KI-Lösungen begann in den 1980er/1990er Jahren in der Form von computer aided detection (CAD)-Systemen [4], die vor allem bei standardisierbaren Untersuchungen, z. B. bei Mammographien [5] oder zur Detektion pulmonaler Rundherde in Computertomographie (CT)- oder Röntgen-Untersuchungen des Thorax eingesetzt wurden [6]. Obgleich insbesondere in den USA zunächst verbreitet angewendet, blieb der Einsatz von CAD umstritten [4]. Verschiedene Studien belegten den eingeschränkten Nutzen oder zeigten sogar, dass sich die diagnostische Performance durch den Einsatz von CAD verschlechtern konnte [7, 8].

Durch die kontinuierliche Zunahme an zu befundendem Bildmaterial bei gleichzeitig ausbleibendem, qualifiziertem Nachwuchs an Radiologen hat das Interesse an automatisierbarer, computerunterstützender Technologie trotz der umstrittenen Ergebnisse von CAD über die Jahre nicht abgenommen. Der aktuelle Clinical Workforce Census des Royal College of Radiologists aus dem Jahr 2024 hat für das Vereinigte Königreich ein aktuelles Defizit von 29 % beim Bedarf an Radiologen berechnet, das innerhalb von fünf Jahren auf 39 % steigen könnte – was einem zusätzlichen Bedarf an 1.953 Fachärzten für Radiologie (bei aktuell 4.699 beschäftigten Fachärzten) entspricht [9]. Durch die bloße Verkürzung der Befundungszeiten kann diesem Trend jedoch nur

[1] Aus Gründen der besseren Lesbarkeit wird im Text die grammatikalisch männliche Form von Personenbezeichnungen verwendet. Sie steht stellvertretend für alle Geschlechter und impliziert keine Wertung. Die Entscheidung dient ausschließlich der sprachlichen Vereinfachung sowie der korrekten grammatikalischen Anwendung der Wortstämme.

unzureichend entgegengewirkt werden, da sich damit einhergehend auch die Fehlerquote erhöht [10]. Hingegen verspricht die Anwendung von KI-Lösungen, die entstehende Lücke zu schließen, indem sie die Befundung beschleunigt, dabei die Befundqualität weiterhin hochhält oder sogar verbessert und sich gleichzeitig reibungslos in den bestehenden klinischen Workflow einbinden lässt.

Im Laufe der Zeit haben sich die zur Verfügung stehenden KI-Lösungen technisch deutlich weiterentwickelt. Zunehmend kommen immer komplexere Lösungen des maschinellen Lernens zum Einsatz. In der Regel finden dabei neuronale Netzwerke Anwendung, die aus einer Eingabeschicht (*input layer*), einer oder mehreren Zwischenschichten (*hidden layer*) und einer Ausgabeschicht (*output layer*) bestehen. Die Schichten umfassen jeweils künstliche Neuronen, die mit den benachbarten Schichten über Aktivierungsfunktionen verknüpft sind [11]. Je mehr Rechenleistung zur Verfügung steht, desto komplexere Netzwerkstrukturen sind möglich, wobei die Komplexität von der Anzahl der Zwischenschichten bestimmt wird. Der in diesem Kontext geläufige Begriff *Deep Learning* beschreibt diesen komplexen Informationsverarbeitungsprozess. Maschinelles „Lernen“ entsteht dadurch, dass neuronale Netzwerke die Aktivierungsfunktionen zwischen den neuronalen Schichten auf Basis der bekannten Trainingsdaten selbst anpassen können. So ist es möglich, ein dediziertes Training hinsichtlich einer bestimmten Fragestellung durchzuführen (z. B. Pathologieerkennung auf radiologischem Bildmaterial).

In den letzten Jahren hat die Anzahl von Deep-Learning-Algorithmen mit Anwendungsmöglichkeit in der Medizin stetig zugenommen. Hochrangig publizierte Studien konnten einzelnen, vortrainierten Algorithmen in der klinischen Anwendung eine sehr hohe Leistungsfähigkeit attestieren. Überzeugende Ergebnisse fanden sich etwa bei der Detektion von Pathologien des Augenhintergrunds [12, 13] oder bei der Differenzierung von suspekten Hautläsionen [14], wobei hier die Algorithmen teilweise sogar eine bessere Performance als medizinische Experten des jeweiligen Fachgebiets zeigten. Auch im Bereich der Radiologie wurden zahlreiche hochrangig publizierte Studien mit exzellenter KI-Performance veröffentlicht – etwa zur Detektion kritischer Befunde in der CT des Schädels [15], bei der Erkennung von Lungenrundherden [16] oder im Rahmen von Screening-Untersuchungen der Mammographie [17, 18, 19]. Die Anzahl an KI-Algorithmen in der Radiologie nimmt stetig zu – mittlerweile existiert bereits eine Vielzahl kommerziell verfügbarer Lösungen [20, 21, 22].

Parallel zur Weiterentwicklung klassischer Deep-Learning-Modelle, die vornehmlich Bilddaten analysieren, gewinnen zunehmend sogenannte Large Language Models (LLMs) an Bedeutung. Aktuelle Studien zu LLMs befassen sich beispielsweise mit der Verbesserung radiologischer Befundberichte [23, 24], der strukturierten Aufarbeitung onkologischer Krankheitsverläufe [25] sowie der kombinierten Evaluation von Bild- und Textmaterial [26]. Der Fokus dieser Arbeit und der beschriebenen Studien liegt primär auf bildbasierten Deep-Learning-Algorithmen, die in der radiologischen Diagnostik weiterhin klassischerweise Anwendung finden.

Die zunehmend leistungsfähige Detektion von Mustern und Pathologien auf radiologischem Bildmaterial wird als vielversprechend diskutiert: Einerseits bietet sie eine potenziell erhöhte diagnostische Qualität (durch die Reduktion übersehener Befunde beim *second reader*-Ansatz und die dadurch möglicherweise verbesserte differentialdiagnostische Einstufung) und damit eine verbesserte Patientenversorgung (z. B. frühere Diagnose und damit auch frühere adäquate Therapieeinleitung). Andererseits kann etwa durch eine Vorbefundung durch den Algorithmus potenziell auch eine Optimierung des Workflows durch Arbeitslisten-Triagierung erreicht werden. Gleichermaßen gilt es zu bedenken, dass der Entscheidungsprozess des Algorithmus (im Bereich der *hidden layer*) nicht dem üblichen radiologischen Vorgehen entspricht und die Nachvollziehbarkeit von KI-Ergebnissen daher nicht immer gegeben ist (Stichwort: *explainable AI* [*artificial intelligence*]). Das daraus resultierende ethische Dilemma wird auch als *black box*-Problem bezeichnet [27]. Dabei gilt es zu berücksichtigen, dass der Entscheidungsprozess der KI maßgeblich von der Qualität der Trainingsdaten abhängt. Sind die Trainingsdaten nicht ideal gewählt, können Confounder auftreten – etwa die fälschliche Detektion einer Thoraxdrainage als Pneumothorax im Röntgenbild [28]. Eine Rolle spielt dabei auch die Art der Annotation (z. B. reine Nennung des Befunds vs. Einzeichnung der Pathologie) [29]. Um den technischen Besonderheiten und dem Umstand gerecht zu werden, dass manche radiologische Modalitäten einer erheblichen Inter-Reader-Variabilität unterliegen (z. B. Röntgen-Thorax [30]), ist es notwendig, präzise klinische Testungen durchzuführen, die die Validität eines Algorithmus belegen (interne und externe Validierung) [31]. Ein robuster Algorithmus muss seinen Zweck unabhängig von Gerät sowie Institut oder Klinik erfüllen können.

In dieser Habilitationsschrift werden verschiedene Deep-Learning-Algorithmen aus den Bereichen Röntgen-Thorax und Magnetresonanztomographie (MRT)-Schädel (Hirnvolumetrie) kli-

nisch evaluiert. Dabei liegt der Fokus klar auf dem klinischen Anwendungsszenario und dem Einsatz im klinischen Alltag – auch unter Bezugnahme auf verschiedene medizinische Berufsgruppen (z. B. Einschluss von Nicht-Radiologen oder Neuroradiologen vs. Radiologen). Zudem wird untersucht, wie sich die Einführung eines bereits mit überzeugenden Ergebnissen extern validierten Algorithmus (zur Frakturdetektion) auf die Wahrnehmung von Radiologen im klinischen Alltag auswirkt und wie ein KI-Algorithmus (Röntgen-Thorax) weiter optimiert werden kann, um beispielsweise den Besonderheiten eines Instituts oder einer Klinik gerecht zu werden.

Im Folgenden werden die Originalarbeiten als kumulative Habilitationsschrift zusammengefasst und in den jeweiligen wissenschaftlichen Kontext eingeordnet.

2 Externe Validierung von KI-Algorithmen

2.1 Externe Validierung eines Röntgen-Thorax-Algorithmus im Notaufnahme-Setting

Artificial Intelligence in Chest Radiography Reporting Accuracy: Added Clinical Value in the Emergency Unit Setting Without 24/7 Radiology Coverage

Rudolph, J., Huemmer, C., Ghesu, F. C., Mansoor, A., Preuhs, A., Fieselmann, A., Fink, N., Dinkel, J., Koliogiannis, V., Schwarze, V., Goller, S., Fischer, M., Jörgens, M., Ben Khaled, N., Vishwanath, R. S., Balachandran, A., Ingrisch, M., Ricke, J., Sabel, B. O., & Rueckel, J. (2022). Artificial Intelligence in Chest Radiography Reporting Accuracy: Added Clinical Value in the Emergency Unit Setting Without 24/7 Radiology Coverage. Investigative radiology, 57(2), 90–98.

https://doi.org/10.1097/RLI.0000000000000813

Journal impact factor (2022): 6.7

Die Röntgenuntersuchung des Thorax ist nach wie vor ein zentraler Baustein in der Primärdiagnostik und wird täglich weltweit in der Notaufnahme eingesetzt [32, 33]. An der Befundung der Bilder beteiligen sich in der Regel nicht nur Radiologen, sondern auch darüber hinaus ärztliches Personal aus anderen Fachdisziplinen (z. B. innere Medizin oder Chirurgie). Kommt es zu einem hohen Workload oder fehlt eine vollumfassende 24/7-Abdeckung durch eine radiologische Abteilung (z. B. in kleineren Krankenhäusern) erfolgt die Befundung evtl. ausschließlich durch Nicht-Radiologen. In der o. g. Originalarbeit wurde ein auf die Detektion von vier spezifischen Pathologien trainierter KI-Algorithmus (pneumonieverdächtige Konsolidierungen, Pleuraergüsse, Pneumothoraces und pulmonale Rundherde) an einer Notaufnahmekohorte mit gezielt angereicherten Pathologien sowie „unauffälligen“ Kontrollfällen extern validiert. Das entstandene Kollektiv, bestehend aus 563 Röntgen-Thoraces (posterior-anterior Projektion, ausschließlich Untersuchungen aus der Notaufnahme), wurde von insgesamt neun verschiedenen Ärzten (drei radiologischen Fachärzten, drei radiologischen Assistenzärzten und drei nicht-radiologischen Assistenzärzte mit Notaufnahmeerfahrung) durch Anwendung eines Likert-skalierten, pathologiespezifischen Konfidenz-Scores jeweils unabhängig ausgewertet. Die Auswertungen der radiologischen Fachärzte dienten dabei als Referenzstandard. Als besondere wissenschaftliche Neuheit ist hierbei einerseits die statistische Berücksichtigung der generellen Unsicherheit in der Röntgen-Thorax-Befundung durch Verwendung verschiedener Konfidenzniveaus und andererseits auch der

Einschluss nicht-radiologischen ärztlichen Personals hervorzuheben.

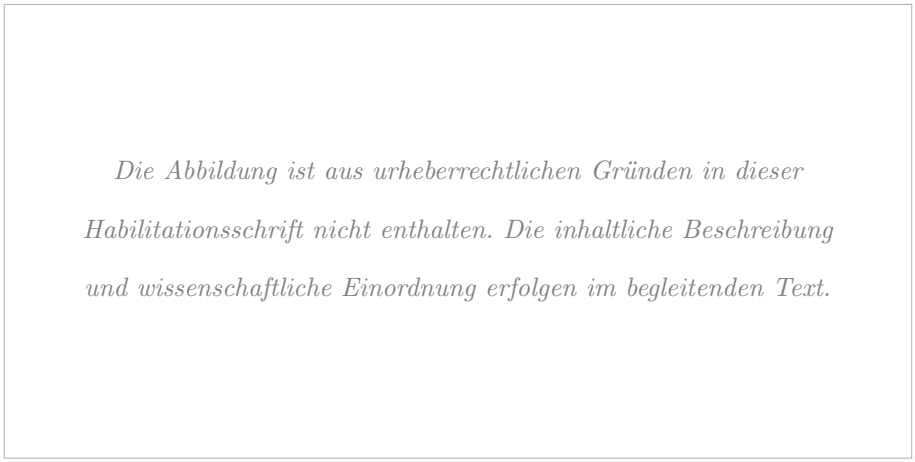

Abbildung 1: Matrix der ROC-Analysen für die Pathologien Pneumothorax (A1-A4), Pleuraerguss (B1-B4) und Pneumonie-suspekte Konsolidierungen (C1-C4), jeweils unter Berücksichtigung der in der Sensitivität aufsteigenden Referenzstandards (RFS I-IV). Abgebildet sind die Performances des KI-Algorithmus, der radiologischen und der nicht-radiologischen Assistenzärzte. Performance-Metriken wurden für die nach Youden-Statistik optimierte Threshold berechnet. Die Abbildung entspricht Figure 4 aus Rudolph et al. [34].

Betrachtet man den sensitivsten Referenzstandard (RFS IV), so konnte der getestete KI-Algorithmus in der receiver operating characteristic (ROC) Auswertung areas under the curve (AUCs) von 0,940 $[0,893-0,988]$ für die Detektion von Pneumothoraces (Abbildung 1 (A4)), 0,953 $[0,933-0,973]$ für die Detektion von Pleuraergüssen (Abbildung 1 (B4)) und 0,858 $[0,813-0,903]$ für die Detektion von pulmonalen Rundherde (Abbildung 2 (A4)) erzielen, was jeweils vergleichbar war mit dem Performance-Niveau der erfahrenen radiologischen Assistenzärzte, die Performance der Nicht-Radiologen aber deutlich überstieg. Bei der Detektion der Pneumonie-suspekten Konsolidierungen schneidet der KI-Algorithmus mit einer AUC von 0,847 $[0,808-0,887]$ (Abbildung 1 (C4)) schlechter ab als die radiologischen Assistenzärzte, aber auf ähnlichem Niveau wie die drei nicht-radiologischen Assistenzärzte.

Aus den Ergebnissen lässt sich schließen, dass der validierte KI-Algorithmus das Potential hat, insbesondere Nicht-Radiologen im Notaufnahme-Szenario zu unterstützen. Ein Ergebnis, das in der folgenden Follow-Up-Studie weiter untersucht wurde (vgl. 2.2).

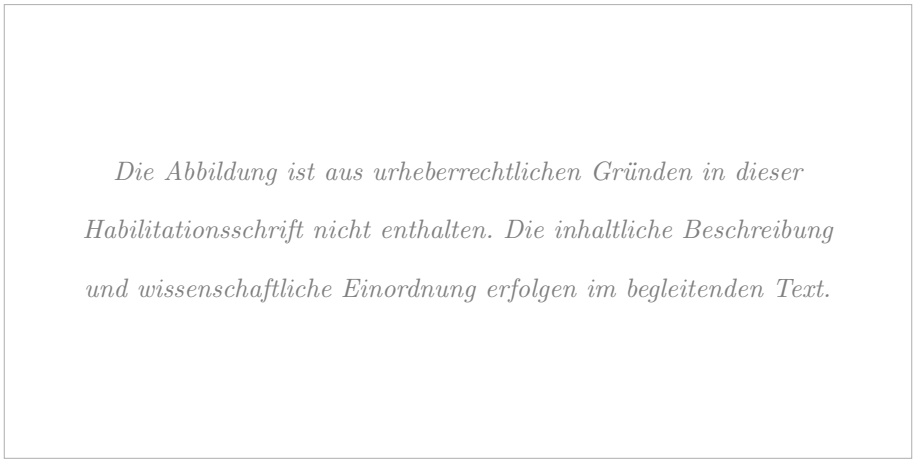

Abbildung 2: ROC-Matrix analog zu Abbildung 1. Dargestellt sind die Analysen für die Pathologie pulmonale Rundherde mit Berücksichtigung aller detektierter Rundherde (A1-A4) bzw. nur unter Berücksichtigung von suspekten Rundherden, bei deren Detektion auch ein CT zur weiteren Abklärung empfohlen wurde (B1-B4). Die Abbildung entspricht Figure 5 aus Rudolph et al. [34].

2.2 Nicht-Radiologen profitieren in der notfalldiagnostischen Röntgen-Thorax-Befundung signifikant von KI-Unterstützung

Nonradiology Health-Care Professionals Significantly Benefit From AI Assistance in Emergency-Related Chest Radiography Interpretation

Rudolph, J., Huemmer, C., Preuhs, A., Buizza, G., Hoppe, B. F., Dinkel, J., Koliogiannis, V., Fink, N., Goller, S. S., Schwarze, V., Mansour, N., Schmidt, V. F., Fischer, M., Jörgens, M., Ben Khaled, N., Liebig, T., Ricke, J., Rueckel, J., & Sabel, B. O. (2024). Nonradiology Health-Care Professionals Significantly Benefit From AI Assistance in Emergency-Related Chest Radiography Interpretation. Chest, 166(1), 157–170.

https://doi.org/10.1016/j.chest.2024.01.039

Journal impact factor (2024): 8.6

Die aufgeführte Originalarbeit ist eine Follow-Up-Studie von 2.1. In dieser Studie wurden dieselbe Kohorte (563 Röntgenbilder aus der Notaufnahme), derselbe KI-Algorithmus zur Röntgen-Thorax-Befundung benutzt sowie dieselben neun Ärzte (drei radiologischen Fachärzte, drei radiologischen Assistenzärzte und drei nicht-radiologischen Assistenzärzte mit Notaufnahmeerfahrung) zur Befundung aufgefordert. Diesmal lagen den Ärzten jedoch neben den originalen Bilddaten auch zusätzliche Bilddaten mit Overlays der Detektionen durch die KI vor (*secondary captures*). Das zweite Reading erfolgte nach einer *washout*-Phase von ca. 12 Monaten. Die Ärzte waren angehalten, ihre Ergebnisse aus dem ersten Reading (ohne direkte KI-Unterstützung, vgl. 2.1) nicht zu betrachten. In der Auswertung zeigte sich, dass die Nicht-Radiologen deutlich von der KI-Assistenz in allen vier getesteten Pathologien profitieren konnten (Pneumonie-suspekte Konsolidierungen, Pleuraergüsse, Pneumothoraces und pulmonale Rundherde). Betrachtet man den sensitivsten Fachärztestandard, so konnten sich die Nicht-Radiologen in der Detektion der zeitkritischen Pathologie Pneumothorax von einer AUC von 0,846 $[0,785 - 0,907]$ ohne KI-Unterstützung auf 0,974 $[0,947 - 1,000]$ mit KI-Unterstützung verbessern ($p < 0,001$). Bei gleichbleibender (optimierter) Spezifität konnte eine Zunahme von 30 % Sensitivität und 2 % Accuracy durch die KI-Unterstützung erreicht werden (Abbildung 3).

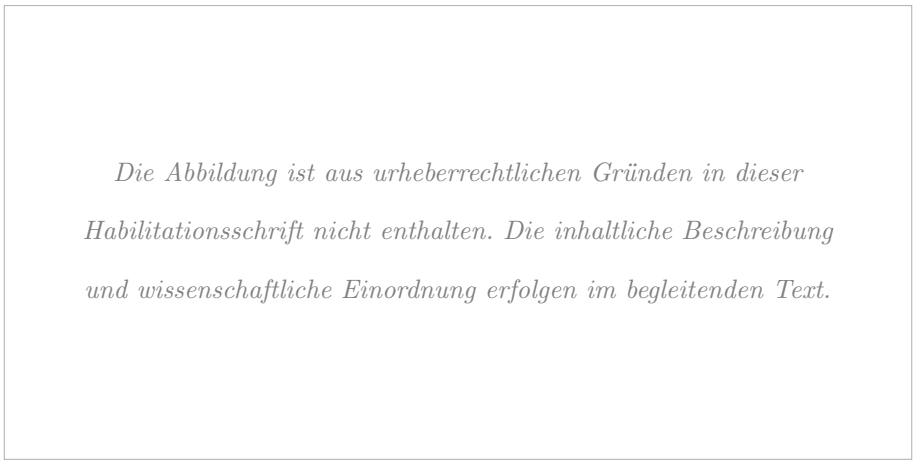

Abbildung 3: Auswertung der Performances für die Pathologie Pneumothorax. Von links nach rechts werden die in der Sensitivität aufsteigenden Referenzstandards (RFS I-IV) gezeigt. Die erste Zeile zeigt die ROC-Auswertungen aus den Quelldaten mit nach Youden-optimierter Threshold und entsprechend berechneten Metriken. Die zweite Zeile zeigt mittels ROC-fitting adaptierte ROC-Kurven, die ein Ablesen der Performance-Verbesserung zwischen erstem Reading (ohne KI-Unterstützung) und zweiten Reading (mit KI-Unterstützung) auf Iso-Spezifitätslinien (vertikale Linien) zulassen. Die dritte Zeile zeigt die Performance-Änderungen von radiologischen Assistenzärzte (RRs) und nicht-radiologischen Assistenzärzte in Sensitivität und Accuracy zwischen erstem und zweitem Reading. Die Abbildung entspricht Figure 2 aus Rudolph et al. [35].

Der größte Unterschied ergab sich in der Detektion von pulmonalen Rundherden. Hier konnten sich die Nicht-Radiologen von einer AUC von 0,723 $[0,661-0,785]$ auf eine AUC von 0,890 $[0,848-0,931]$ ($p<0,001$) mit KI-Unterstützung verbessern, was einer Zunahme der Sensitivität um 53 % und der Accuracy um 7 % (bei jeweils gleichbleibender Spezifität) entspricht (Abbildung 4A).

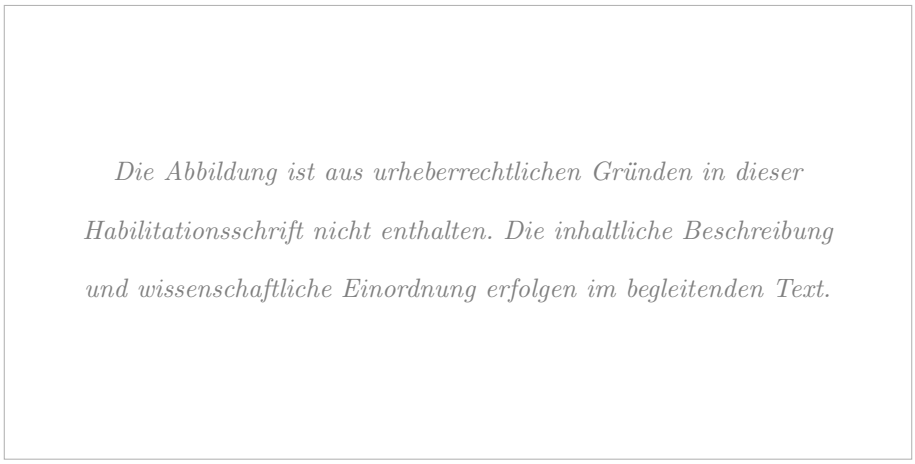

Abbildung 4: Performance-Auswertung für die Pathologie pulmonale Rundherde, analog zu Abbildung 3. (A) Berücksichtigung aller von den Referenzreadern detektierten Rundherde, (B) Berücksichtigung nur der Rundherde, die aus Sicht der Reader eine Follow-up CT gerechtfertigen. Die Abbildung entspricht Figure 5 aus Rudolph et al. [35].

Auch für die Pathologien Pneumonie-suspekte Konsolidierungen und Pleuraergüsse zeigte sich

bei den Nicht-Radiologen ein statistisch signifikanter Performancezugewinn (jeweils $p < 0,001$; vgl. Abbildungen 5 und 6). Die radiologischen Assistenzärzte zeigten kleinere, meistens nicht statistisch signifikante Veränderungen in Performance, Sensitivität und Accuracy.

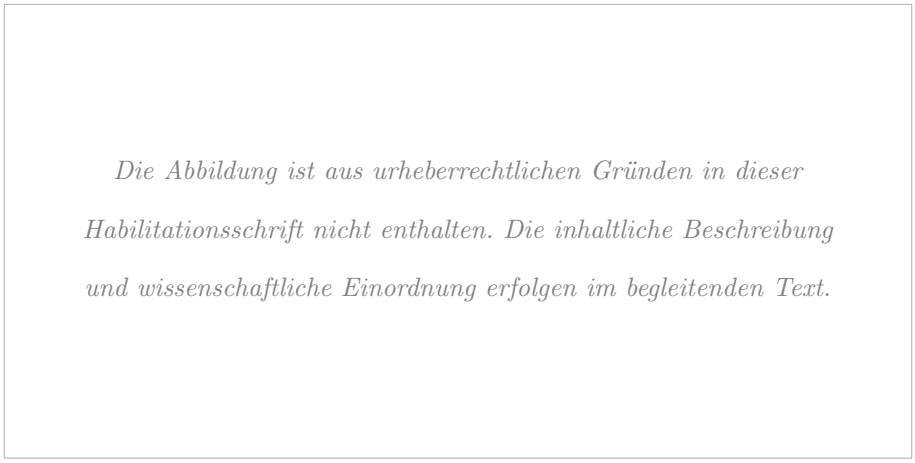

Abbildung 5: Performance-Auswertung für die Pathologie Pleuraerguss, analog zu Abbildung 3. Die Abbildung entspricht Figure 3 aus Rudolph et al. [35].

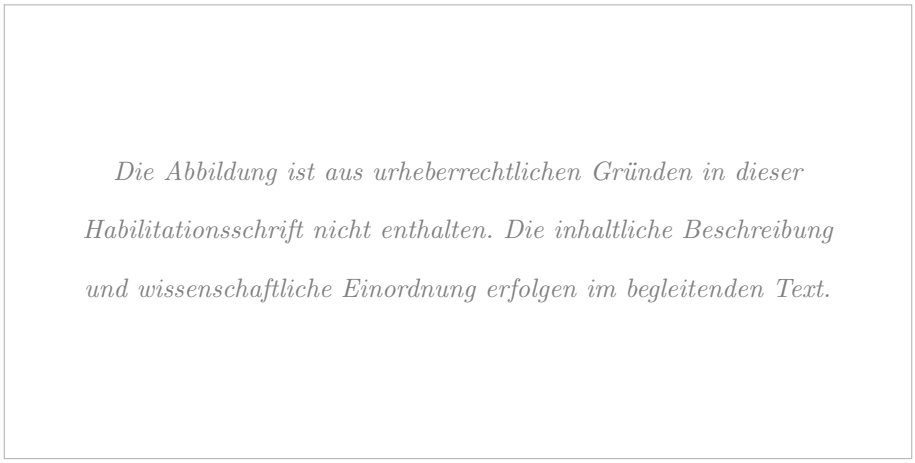

Abbildung 6: Performance-Auswertung für die Pathologie Pneumonie-suspekte Konsolidierung, analog zu Abbildung 3. Die Abbildung entspricht Figure 4 aus Rudolph et al. [35].

Die Ergebnisse verdeutlichen, dass sich der getestete KI-Algorithmus als gutes Unterstützungstool für Nicht-Radiologen in der Notaufnahme eignet und so durch den Zugewinn an Performance eine verbesserte Patientenversorgung potentiell begünstigen kann. Dies ist vor allem dann relevant, wenn ein Krankenhaus keine radiologische Abteilung besitzt oder die radiologische Abteilung nicht durchgehend zur Verfügung steht (hohe Befundlast) bzw. nicht durchgängig besetzt ist.

2.3 Optimiertes Benchmarking von Röntgen-Thorax-Algorithmen anhand mehrerer klinischer Kohorten

Clinically focused multi-cohort benchmarking as a tool for external validation of artificial intelligence algorithm performance in basic chest radiography analysis

Rudolph, J., Schachtner, B., Fink, N., Koliogiannis, V., Schwarze, V., Goller, S., Trappmann, L., Hoppe, B. F., Mansour, N., Fischer, M., Ben Khaled, N., Jörgens, M., Dinkel, J., Kunz, W. G., Ricke, J., Ingrisch, M., Sabel, B. O., & Rueckel, J. (2022). Clinically focused multi-cohort benchmarking as a tool for external validation of artificial intelligence algorithm performance in basic chest radiography analysis. Scientific reports, 12(1), 12764.

https://doi.org/10.1038/s41598-022-16514-7

Journal impact factor (2022): 4.6

In dieser Originalarbeit werden drei klinisch relevante Kohorten zur Evaluation von KI-Algorithmen in der Röntgen-Thorax-Bildgebung vorgestellt. Aus den drei Kohorten ergibt sich eine Benchmarking-Pipeline, die KI-Algorithmen umfassend auf verschiedene Kriterien bzw. klinische Anforderungen prüfen kann. Die Kohorten unterscheiden sich dabei in der Aufnahmeakquisition (liegend vs. stehend Röntgen), im Referenzstandard (Röntgen vs. CT basiert) und der Möglichkeit die Algorithmus-Performance auch mit der Performance nicht-radiologischen ärztlichen Personals zu vergleichen. Die Pipeline umfasst:

1. die in 2.1 und 2.2 aufgezeigte Kohorte mit 563 Röntgenbildern aus der Notaufnahme (posterior-anterior; Evaluation durch 9 Ärzte, darunter Radiologen und Nicht-Radiologen; vier Pathologien, vgl. 2.1 und 2.2),

2. eine Kohorte mit 6.248 Röntgenbildern im Liegen (anterior-posterior), die radiologisch hinsichtlich des Vorliegens eines Pneumothorax, seiner Größe und dem Vorliegen einer einliegenden Thoraxdrainage annotiert wurde[2] und

3. eine 166 Patienten umfassende Kohorte mit Röntgenbildern im Liegen (anterior-posterior), die radiologisch hinsichtlich des Ursprungs dort abgebildeter basaler Konsolidierungen beurteilt wurde (für alle eingeschlossenen Röntgenbilder lag eine zeitnahe korrelierende CT zum Vergleich vor [Zeitfenster beider Akquisitionen $<$ 90 Minuten])[3].

[2] Detailierte Informationen zur Kohorte finden sich in Rueckel et al. [28]
[3] Detailierte Informationen zur Kohorte finden sich in Kunz et al. [36]

Der getestete Algorithmus CheXNet (von Rajpurkar et al. [37, 38], Implementierung von arnoweng von github.com, NLP-basiert, öffentliche Trainingsdaten [39], frei verfügbar) zeigte eine bessere Performance in der Detektion von pulmonalen Rundherden als der Konsensus der radiologischen Assistenzärzte in Kohorte (1) (AUC KI/Konsensus 0,851/0,839; vgl. Abbildung 7 (B4)).

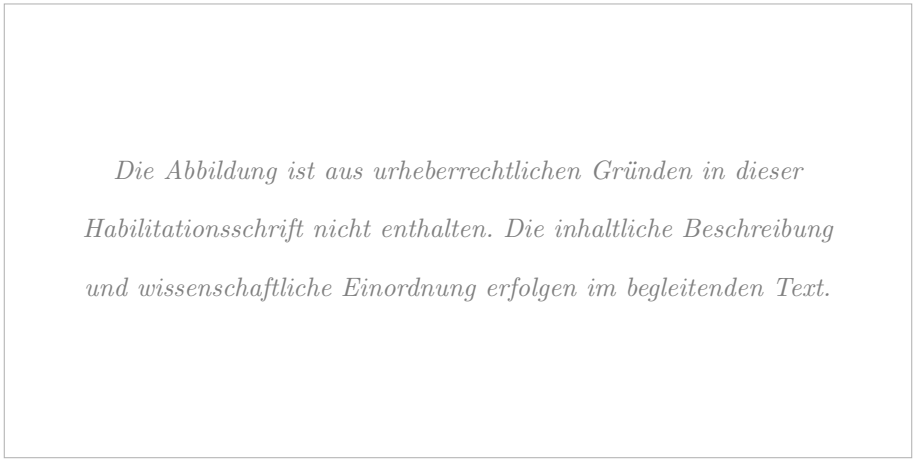

Abbildung 7: KI-Performance-Auswertungen für die Pathologie pulmonale Rundherde. Von links nach rechts werden die Auswertungen für die in der Sensitivität ansteigenden Referenzstandards (RFS I-IV) gezeigt. Die obere Zeile berücksichtigt alle von den Referenzreadern detektierten Lungenrundherde, die untere Zeile berücksichtigt nur die Rundherde, die eine weitere CT-Abklärung gerechtfertigen würden. Verglichen werden die Leistungen verschiedener Algorithmus-Klassifikatoren mit den Leistungen der radiologischen und nicht-radiologischen Assistenzärzte (Kohorte 1). Die Abbildung entspricht Figure 4 aus Rudolph et al. [40].

Zudem schnitt der Algorithmus besser als der radiologische Konsensus aus Kohorte (3) bei der Detektion basaler Pneumonien (AUC KI/Konsensus: 0,825/0,782; vgl. Abbildung 8 [3](A)) und der Detektion von basalen Pleuraergüssen (AUC KI/Konsensus: 0,762/0,710; vgl. Abbildung 8 [3](B)) ab.

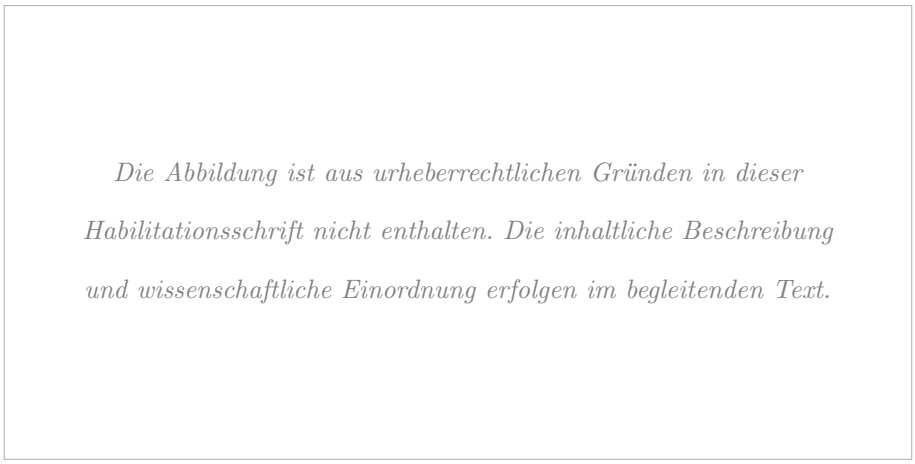

Abbildung 8: KI-Performance-Auswertung für die Pathologien Pleuraerguss und Pneumonie-suspekte Konsolidierung für die entsprechenden Klassifikatoren. [1] und [2] zeigen dabei jeweils die Performance-Auswertungen aus Kohorte 1 im Vergleich zu den Performances von radiologischen und nichtradiologischen Assistenzärzten. [3] zeigt die Performance in Kohorte 3. Die Abbildung entspricht Figure 3 aus Rudolph et al. [40].

Dabei waren die AUC-Werte teilweise höher als in der Originalarbeit angegeben (nodule: 0,780; infiltration: 0,735; effusion: 0,868; vgl. [37]). Der Algorithmus-eigene Klassifikator „infiltration“ zeigte zudem eine starke Abhängigkeit von der Patientenlagerung (besser beim Röntgen im Stehen, schlechter beim Röntgen im Liegen). Die Pneumothorax-Detektion unterlag einer schlechten Performance, wenn keine Thoraxdrainagen auf dem Bild zu erkennen waren (Kohorte 2). Dies deutet auf einen systematischen Fehler im Trainingsprozess des Algorithmus hin.

Die vorgestellte Pipeline bietet ein umfangreiches Benchmarking von Röntgen-Thorax-Algorithmen mit Berücksichtigung potentieller Confounder, unterschiedlicher Referenzstandards, verschiedener Lagerungen und der Möglichkeit, KI-Ergebnisse mit der Performance unterschiedlich qualifizierter Ärzte zu vergleichen.

2.4 Validierung eines Algorithmus zur Detektion von Endotrachealtuben und zentralvenösen Kathetern

Artificial Intelligence to Assess Tracheal Tubes and Central Venous Catheters in Chest Radiographs Using an Algorithmic Approach With Adjustable Positioning Definitions

Rueckel, J., Huemmer, C., Shahidi, C., Buizza, G., Hoppe, B. F., Liebig, T., Ricke, J., **Rudolph, J. (geteilte Letztautorschaft)***, & Sabel, B. O.* (2024). Artificial Intelligence to Assess Tracheal Tubes and Central Venous Catheters in Chest Radiographs Using an Algorithmic Approach With Adjustable Positioning Definitions. Investigative radiology, 59(4), 306–313.

https://doi.org/10.1097/RLI.0000000000001018

Journal impact factor (2024): 8.0

Eigenanteil: Studiendesign und Konzeption, Unterstützung bei der Auswertung, Erstellung und Überarbeitung des Manuskripts/Revision

Diese Originalarbeit beschreibt und validiert einen KI-Algorithmus, der die Lage von zentralvenösen Kathetern (ZVKs) und Endotrachealtuben auf Röntgen-Thorax-Aufnahmen beurteilt. ZVKs und Endotrachealtuben werden insbesondere in der Notfallmedizin häufig angelegt. Die korrekte Lage ist essentiell für die Patientensicherheit und das erwünschte Outcome und wird in der Regel durch ein Röntgenbild des Thorax bestätigt [41, 42, 43].

Der vorgestellte Algorithmus evaluiert die Qualität der Lage anhand der räumlichen Lagebeziehungen der Spitzen von ZVKs/Tuben zu erkennbaren anatomischen Strukturen. Bei der Analyse der ZVKs wird eine *region of interest* definiert, die durch Segmentierung anatomischer Landmarken die korrekte Lage eingrenzen soll. Über die Einführung eines *multitask neural network* ermöglicht der Algorithmus die gleichzeitige Klassifizierung der ZVKs/Tuben (Typ/Existenz) sowie eine Verlaufssegmentierung und Erkennung der Katheter- und Tubenspitzen.

Die Validierung erfolgte auf insgesamt 589 Röntgen-Thorax-Aufnahmen im Liegen, welche radiologisch hinsichtlich des Vorliegens von ZVKs/Tuben und der Einschätzung, ob eine korrekte Positionierung vorliegt, annotiert wurden. Die Erkennung der korrekten/inkorrekten Fremdmateriallage durch den Algorithmus wurde mithilfe von ROC-Analysen beurteilt. In einer zweiten Evaluation (Reading 2) nach einer mehrwöchigen *washout*-Phase konnten mithilfe eines Software-

Tools Bildpositionen der vom Algorithmus erkannten ZVKs/Tubusspitzen korrigiert werden, wodurch die Lokalisierungsgenauigkeit quantifiziert werden konnte.

ZVKs wurden in 98 % (Abbildung 9A) und Endotrachealtuben in 100 % der Fälle (Abbildung 10A) auf Thoraxliegendaufnahmen durch den Algorithmus korrekt klassifiziert.

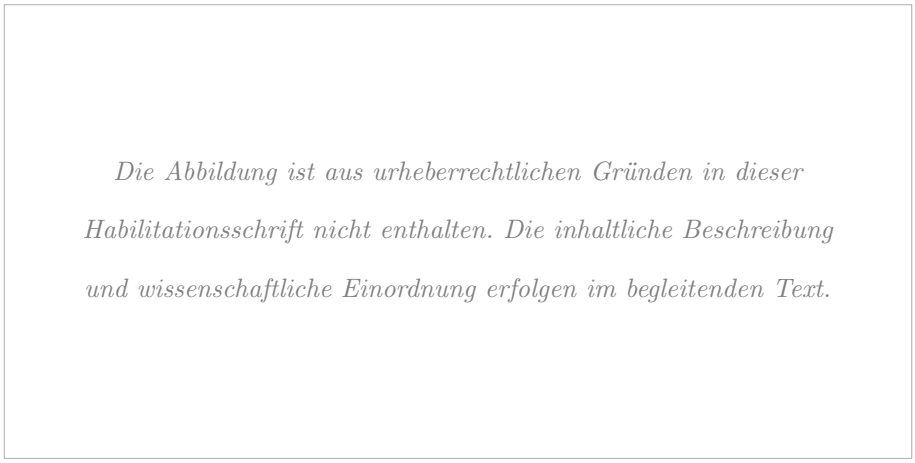

Abbildung 9: Bewertung der räumlichen Erkennungsgenauigkeit von ZVKs. Der geringe Anteil an Diagonaleinträgen in der Konfusionsmatrix (A) zeigt die hohe Erkennungsgenauigkeit von ZVKs. Insgesamt wurden 231 von 588 Katheterspitzen (39,3 %) in der zweiten Auswertung korrigiert, was sich auf 34 % der vom Annotator geänderten Bilder bezieht – darunter 13 Bilder mit hinzugefügten oder entfernten Katheterspitzen und 188 Bilder mit modifizierten Katheterspitzen. 77 % aller Korrekturen betrugen weniger als 3 mm in vertikaler sowie horizontaler Richtung (B). (C) zeigt eine histogrammbasierte Darstellung von Ausreißern mit großen Korrekturen, die manuell überprüft wurden. Die Abbildung entspricht Figure 5 aus Rueckel et al. [44].

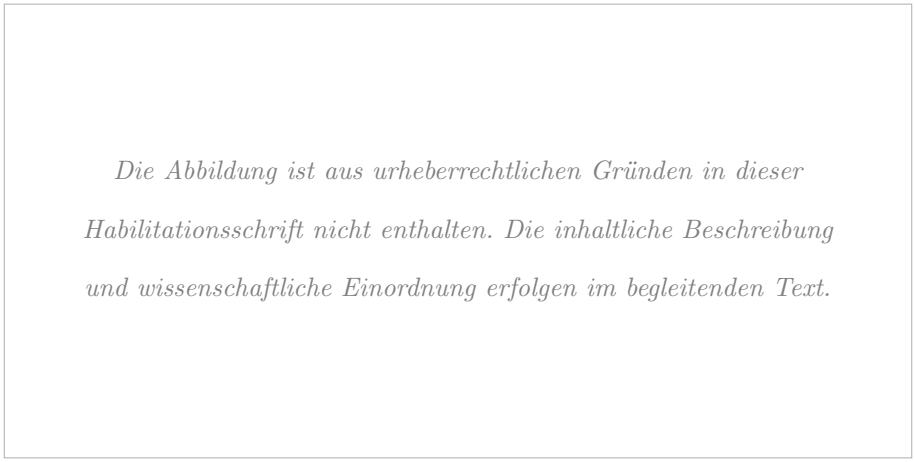

Abbildung 10: Bewertung der räumlichen Erkennungsgenauigkeit von Endotrachealtuben. (A) Das Fehlen von Diagonaleinträgen in der Konfusionsmatrix zeigt, dass alle Tuben vom Algorithmus erkannt wurden. (B) 101 von 351 Tubusspitzenpositionen wurden in der zweiten Auswertung korrigiert, wobei die vertikale Korrektur in über 86 % der Fälle weniger als 3 mm betrug. (C) Histogrammbasierte Darstellung der Benutzerkorrekturen. Ausreißer mit großen Korrekturen, die manuell überprüft wurden, sind gekennzeichnet. (D) Das Streudiagramm zeigt einen kleinen systematischen Bias bei den vertikalen Korrekturen, die überwiegend in kaudaler Richtung vorgenommen wurden (90 Fälle). Ein systematischer Bias in horizontaler Richtung wurde nicht beobachtet. Die Abbildung entspricht Figure 3 aus Rueckel et al. [44].

Die Erkennung der Katheter- und Tubenspitzen erfolgte mit hoher Genauigkeit: Bei ZVKs wurde die Annotation der Spitze in 77 % um weniger als 3 mm nachträglich korrigiert (Abbildung 9B) und bei Endotrachealtuben in > 86 % (Abbildung 10B). Die AUC bei der Erkennung falsch positionierter Devices betrug bei ZVKs > 0,96 (ZVKs-Spitze in falschem Gefäß; vgl. Abbildung 11A) bzw. > 0,93 (auch suboptimale Lage berücksichtigt; vgl. Abbildung 11B) und bei Endotrachealtuben > 0,98 (Abbildung 12). Einschränkungen bei der Klassifikation einer suboptimalen ZVKs-Lage ergaben sich dabei hauptsächlich durch Limitationen in der verwendeten Lagedefinition (*region of interest* anhand von anatomischen Landmarken).

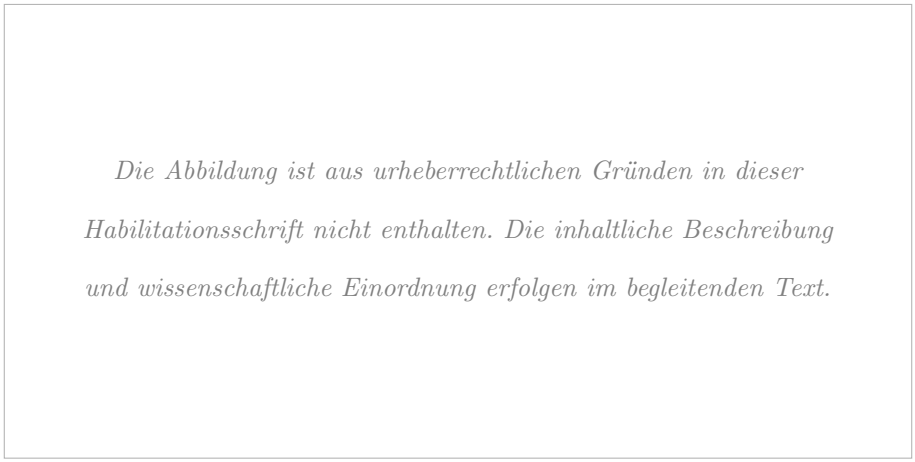

Abbildung 11: ZVKs (ROC-Analyse). (A)–(C) zeigen die ROC-Kurven der KI (orange) und der korrigierten Ergebnisse aus Reading 2 (blau), verglichen mit verschiedenen Referenzstandards. (A) nur klare Fehlplatzierungen, Referenzstandard 1a). (B) Fehlplatzierung + suboptimale Länge, Referenzstandard 1b). (C) Präzise Spitzen- und ROI-Annotationen, Referenzstandard 2. (D) Grundlage der ROC-Berechnung: Abstand der Katheterspitze zum ROI-Zentrum. Die Abbildung entspricht Figure 6 aus Rueckel et al. [44].

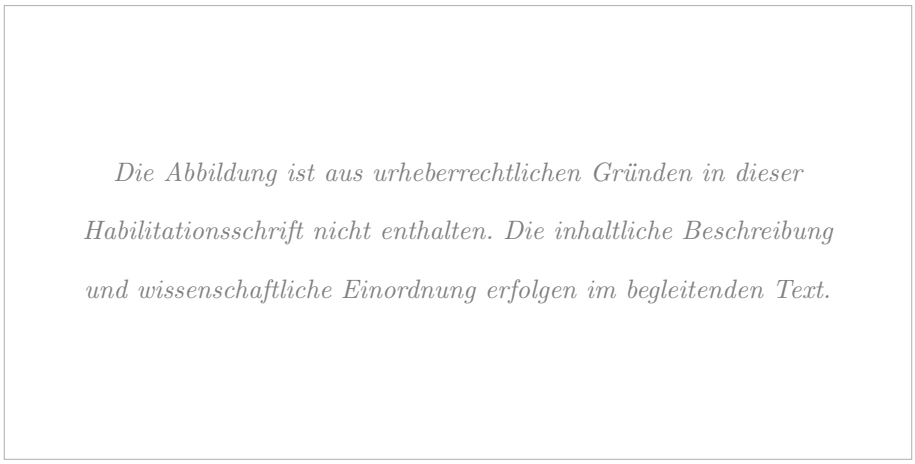

Abbildung 12: Endotrachealtuben (ROC-Analyse). (A)–(C) zeigen die ROC-Kurven der KI allein (orange) und der korrigierten Erkennungen aus Reading 2 (blau), jeweils im Vergleich zu: (A) Schwelle 5mm (nur klare Fehlplatzierungen, Referenzstandard 1a); (B) Schwelle 20 mm (Fehlplatzierung + suboptimale Lage), Referenzstandard 1b); (C) exakten Annotationen von Tubusspitze und Carina (Referenzstandard 2). Die Abbildung entspricht Figure 4 aus Rueckel et al. [44].

Die Studie zeigt, dass ZVKs und Endotrachealtuben durch den vorgestellten Algorithmus präzise lokalisiert werden können. Die klinische Anwendung zur schnellen Triagierung (korrekte Lage vs. Fehllage) ist jedoch durch die vage Definition der suboptimalen Positionierung von ZVKs limitiert. Der Algorithmus könnte jedoch im Bedarfsfall an spezifische Benutzer- oder Patientengruppenanforderungen angepasst werden.

2.5 KI-gestützte Hirnvolumetrie in der MRT des Schädels

Artificial Intelligence Based Rapid Brain Volumetry Substantially Improves Differential Diagnosis in Dementia

Rudolph, J., Rueckel, J., Döpfert, J., Ling, W., Opalka, J., Brem, C., Hesse, N., Ingenerf, M., Koliogiannis, V., Solyanik, O., Hoppe, B.F., Zimmermann, H., Flatz, W., Forbrig, R., Patzig, M., Rauchmann, B.-S., Perneczky, R., Peters, O., Priller, J.,Schneider, A., Fliessbach, K., Hermann, A., Wiltfang, J., Jessen, F., Düzel, E., Bürger, K., Teipel, S., Laske, C., Synofzik, M., Spottke, A., Ewers, M., Dechent, P., Haynes, J.-D., Levin, J., Liebig, T., Ricke, J., Ingrisch, M. & Stoecklein, S. (2024). Artificial intelligence-based rapid brain volumetry substantially improves differential diagnosis in dementia. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 16(4), e70037.

https://doi.org/10.1002/dad2.70037

Journal impact factor (2024): 4.4

In dieser Originalarbeit wird ein KI-Algorithmus zur automatisierten Volumetrie des Gehirns in der MRT vorgestellt und mit dem klinischen Fokus „Differentialdiagnostik der Demenzerkrankungen“ extern validiert. Der Algorithmus zeichnet sich gegenüber anderen, vorwiegend wissenschaftlich verwendeten Lösungen [45, 46, 47, 48] dadurch aus, dass er (1) bei Vorliegen entsprechender Hardware eine sehr schnelle Segmentierung vornimmt ($<$ 5 min), (2) eine vollständige Integration in das bestehende Picture Archiving and Communication System (PACS) ermöglicht und (3) die Ergebnisse der Segmentierungen mit alters- und geschlechtsadaptierten Normwerten vergleichen kann. Die Lösung ist somit im klinischen Alltag nutzbar. Im Rahmen der KI-Validierung wurde ein Kollektiv mit sehr gut klinisch validierten Bilddaten (3D T1-gewichtete MRT-Sequenzen) von Studienpatienten des Deutschen Zentrums für Neurodegenerative Erkrankungen (DZNE) verwendet (Kohorten: DELCODE, DESCRIBE-FTD und DANCER). Insgesamt wurden 55 Patienten eingeschlossen, darunter 17 Patienten mit der Diagnose *Morbus Alzheimer*, 18 Patienten mit der Diagnose *Frontotemporale Demenz* (12 bvFTD, 5 PFNA und 1 SemD) und 20 *gesunde Kontrollpatienten ohne Diagnose einer Demenzerkrankung.* Insgesamt sieben Diagnostiker – darunter zwei langjährig erfahrene Radiologen mit der Schwerpunktbezeichnung Neuroradiologie, zwei radiologische Fachärzte und drei radiologische Assistenzärzte – evaluierten die 3D T1-gewichteten MRT-Sequenzen insgesamt zweimal: (1) Im ersten Reading wurden den Diagnostikern nur die Bilddaten zur Verfügung gestellt. (2) Das zweite Reading erfolgte nach

einer *washout*-Phase von ca. einem Monat. Nun wurden den Diagnostiker neben den Bilddaten zusätzlich KI-erstellte „Volumetrie-Reports“ zur Verfügung gestellt, die eine dedizierte Volumetrie für jeden Hirnlappen und zusätzlich einen Vergleich des gemessenen Volumens mit der alters- und geschlechtsgematchten Vergleichskohorte beinhaltete. In beiden Readings musste eine differentialdiagnostische Abwägung für die drei Diagnosen „Morbus Alzheimer“, „Frontotemporale Demenz“ und „gesunde Kontrolle“ erfolgen. Dazu wurden insgesamt 5 *Diagnosepunkte* auf die drei genannten Diagnosen verteilt.

Der Konsensus aller teilnehmenden Diagnostiker zeigte eine signifikant bessere Performance in der differentialdiagnostisch korrekten Detektion der Patienten mit der Diagnose „Morbus Alzheimer“ sofern ein KI-erstellter „Volumetrie-Report“ vorlag (AUC: ohne KI: 0,800, mit KI: 0,926, $p < 0{,}05$; vgl. Abbildung 13). Bei diesen Patienten nahm auch die Anzahl der korrekt zugeordneten Diagnosepunkte signifikant zu ($p < 0{,}01$; vgl. Abbildung 14) und die Anzahl der falsch zugeordneten Diagnosepunkte signifikant ab ($p < 0{,}03$; vgl. Abbildung 14). Die größte Performancesteigerung zeigte sich bei den radiologischen Fach- ($p < 0{,}04$; vgl. Abbildung 14) und Assistenzärzte ($p < 0{,}02$; vgl. Abbildung 14).

Die Abbildung ist aus urheberrechtlichen Gründen in dieser Habilitationsschrift nicht enthalten. Die inhaltliche Beschreibung und wissenschaftliche Einordnung erfolgen im begleitenden Text.

Abbildung 13: ROC-Analyse zur Performance in der radiologischen Differentialdiagnostik von Morbus Alzheimer – ohne (A) und mit KI-Unterstützung (B), sowie Konsensvergleich (C). Abkürzungen: acc = Genauigkeit, AI = Künstliche Intelligenz, AUC = Fläche unter der ROC-Kurve, BCR = Facharzt für Radiologie, BCNR = Schwerpunktfacharzt Neuroradiologie, RR = Assistenzarzt für Radiologie, OC = Gesamtkonsens, sens = Sensitivität, spec = Spezifität, ppv/npv = pos./neg. präd. Wert, fpr/fnr = Falsch-positiv/-negativ-Rate, ctl = „closest top left“. Die Abbildung entspricht Figure 2 aus Rudolph et al. [49].

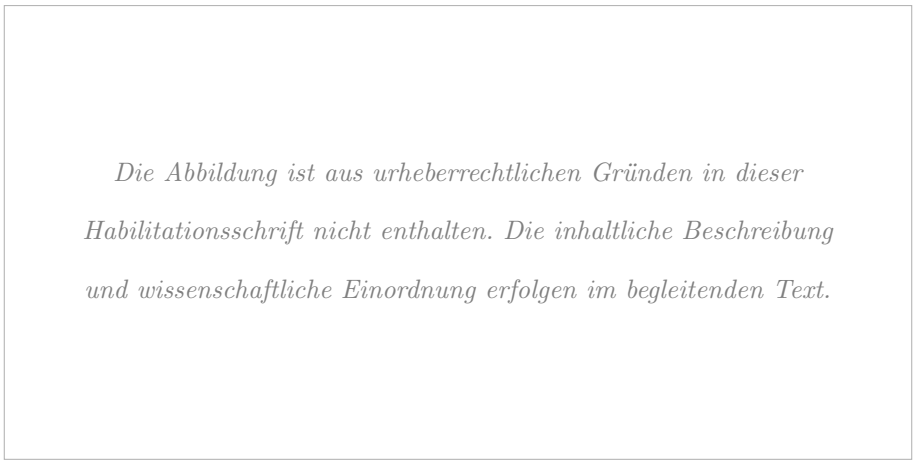

Abbildung 14: Individuelle und konsensusbasierte Performance auf Basis der vergebenen Punkte bei der radiologischen Differentialdiagnostik von Morbus Alzheimer. Abkürzungen analog zu Abbildung 13. Die Abbildung entspricht Table 2 aus Rudolph et al. [49].

Für die Differentialdiagnose „Frontotemporale Demenz“ konnten die korrekt verteilten Diagnosepunkte beim Gesamtkonsensus ($p < 0{,}01$; vgl. Abbildung 15), bei den Neuroradiologen ($p < 0{,}02$; ; vgl. Abbildung 15) und bei den radiologischen Fachärzten ($p < 0{,}05$; ; vgl. Abbildung 15) durch die zusätzlich vorliegenden KI-Auswertungen gesteigert werden.

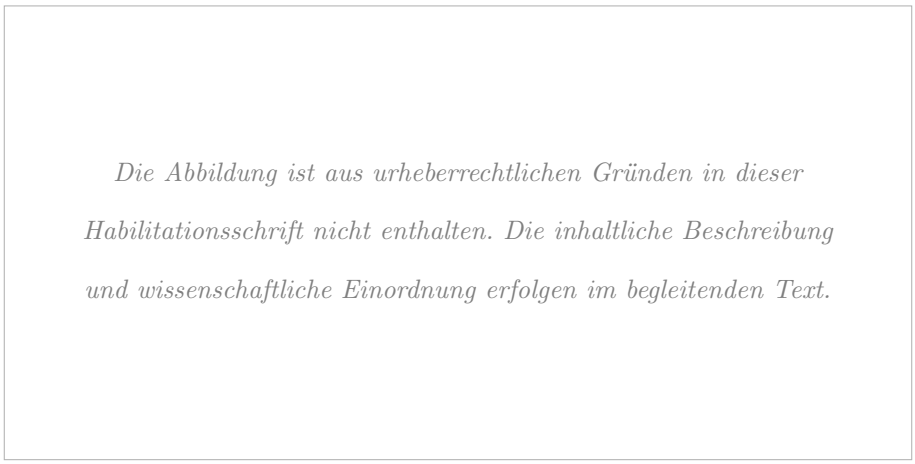

Abbildung 15: Individuelle und konsensusbasierte Performance auf Basis der vergebenen Punkte bei der radiologischen Differentialdiagnostik der Frontotemporalen Demenz. Analog zu Abbildung 14. Die Abbildung entspricht Table 3 aus Rudolph et al. [49].

Die Originalarbeit zeigt, dass alle teilnehmenden Diagnostiker (auch erfahrene Neuroradiologen) von der KI-gestützten Hirnvolumetrie und dem Vergleich der Volumetriemessungen mit alters- und geschlechtsadaptierten Vergleichskollektiven in der Differentialdiagnostik Morbus Alzheimer vs. Frontotemporale Demenz vs. gesund profitieren können.

3 KI in der klinischen Routine

3.1 Wahrnehmung von KI nach Einführung in den klinischen Alltag

Implementing Artificial Intelligence for Emergency Radiology Impacts Physicians' Knowledge and Perception: A Prospective Pre- and Post-Analysis

Hoppe, B. F., Rueckel, J., Dikhtyar, Y., Heimer, M., Fink, N., Sabel, B. O., Ricke, J., **Rudolph, J. (geteilte Letztautorschaft)***, & Cyran, C. C.* (2024). Implementing Artificial Intelligence for Emergency Radiology Impacts Physicians' Knowledge and Perception: A Prospective Pre- and Post-Analysis. Investigative radiology, 59(5), 404–412.

https://doi.org/10.1097/RLI.0000000000001034

Journal impact factor (2024): 8.0

Eigenanteil: Studiendesign und Konzeption, Projektbetreuung/Supervision, Datenerhebung, statistische Analysen einschließlich graphischer Darstellung, Erstellung des ersten Manuskipt-Drafts, Überarbeitung des Manuskripts/Revision

Obgleich die Anzahl an kommerziell verfügbaren KI-Algorithmen in der Radiologie auch zu diesem Zeitpunkt bereits sehr hoch war [20], war die Integration in die klinische Routine Ende 2022 wenig verbreitet. Im Rahmen dieser Originalarbeit wurde untersucht, wie sich die Implementierung eines kommerziell verfügbaren KI-Algorithmus zur Frakturdetektion auf Röntgenbildern in den klinischen Alltag der Notaufnahme, auf die Wahrnehmung von KI und das Wissen über KI-Lösungen bei Radiologen und Klinikern auswirkt. Der getestete Algorithmus wurde vorab ausführlich mit überzeugenden Ergebnissen in externen Validierungsstudien evaluiert [50, 51]. Die Studie wurde als prospektive Interventionsstudie angelegt. Es erfolgte eine zweistufige Umfrage - (1) vor Implementierung des KI-Algorithmus und (2) erneut drei Monate nach Einführung des Algorithmus in die klinische Routine. Zu beiden Zeitpunkten wurden jeweils dieselben Fragen über Erfahrungen/Wissen zu KI und Einschätzung von KI-Lösungen gestellt. Bei der zweiten Umfrage wurde der Fragebogen zudem um Fragen zu den neu erworbenen Erfahrungen mit KI erweitert. Befragt wurden Radiologen und Traumatologen (Unfallchirurgen und Orthopäden). Die Bewertungen der Fragen erfolgten durch Anwendung einer 7-stufigen Likert-Skala (von *-3: stimme absolut nicht* zu bis *+3: stimme absolut zu*). Ein prospektives prä-/post-interventionelles Matching konnte durch selbsterstellte Pseudonymisierungscodes erreicht werden.

Insgesamt nahmen 71 Ärzte an der Umfrage teil, wovon 47 sowohl den prä-interventionellen als auch den post-interventionellen Fragebogen ausfüllten (Follow-up rate: 66 %), darunter 34 Radiologen und 13 Traumatologen. Post-Intervention bewerteten die Teilnehmer die Frage: „KI reduziert übersehene Befunde“ zunehmend positiver (prä-: 1,28 und post: 1,94, $p < 0{,}01$; vgl. Abbildungen 16 und 17). Zudem nahm das Empfinden dafür, dass Ärzte durch KI „sicherer“ werden, signifikant zu (prä: 1,21 und post: 1,64, $p < 0{,}05$; vgl. Abbildungen 16 und 17). Keine signifikante Zunahme gab es hingegen für die Wahrnehmung: „KI macht schneller“ (prä: 0,98, post: 1,21, $p = 0{,}261$; vgl. Abbildungen 16 und 17). Die Frage nach: „KI ersetzt den radiologischen Befund“ wurde nach der Intervention zunehmend negativ bewertet (prä: -2,04, post: -2,34, $p < 0{,}05$; vgl. Abbildungen 16 und 17). Zudem konnte ein edukativer Aspekt nachgewiesen werden: Die Fragen nach der subjektiven „Erfahrung mit klinischen KI-Lösungen“, nach der subjektiven „Einschätzung von Chancen durch KI-Lösungen“ und nach der subjektiven „Einschätzung möglicher Risiken“ wurden nach der Intervention positiver bewertet (in allen Fällen $p < 0{,}01$; vgl. Abbildungen 16 und 17).

Die Abbildung ist aus urheberrechtlichen Gründen in dieser Habilitationsschrift nicht enthalten. Die inhaltliche Beschreibung und wissenschaftliche Einordnung erfolgen im begleitenden Text.

Abbildung 16: Evaluation der Likert-skalierten Bewertungen prä- und post-interventionell. Die Daten werden als Mittelwert ± SD dargestellt. Antworten wurden auf einer 7-Punkte-Likert-Skala von 3 („stimme überhaupt nicht zu“) bis +3 („stimme voll zu“) erfasst. Die Analyse erfolgte an gepaarten Messungen (vor/nach Intervention, $n = 47$) mit dem Wilcoxon-Vorzeichen-Rang-Test. Signifikanz wird durch Sterne angezeigt: *$P < 0{,}05$, **$P < 0{,}01$, ***$P < 0{,}001$; ns = nicht signifikant. Die Abbildung entspricht Table 2 aus Hoppe et al. [52].

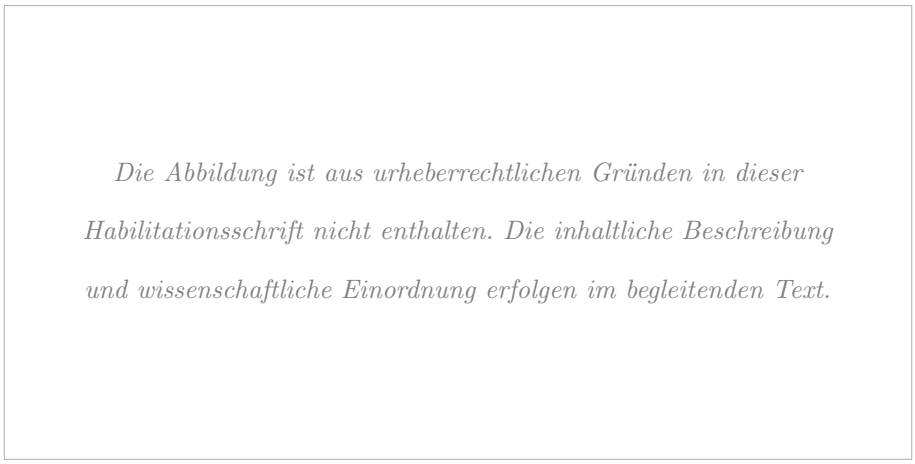

Abbildung 17: Vergleich der Likert-skalierten Bewertungen vor und nach der Intervention. Die Mittelwerte mit Standardabweichung (SD) sind als Fehlerbalken dargestellt. Die Abbildung entspricht Figure 4 aus Hoppe et al. [52].

Die Radiologen nutzten den KI-Algorithmus häufiger als die Traumatologen ($p < 0{,}001$; vgl. Abbildung 18) und schätzten die Vorteile eher höher ein (jeweils $p < 0{,}05$; vgl. Abbildung vgl. Abbildung 18). Je älter der Teilnehmer desto weniger wahrscheinlich wurden KI-Lösungen genutzt oder positiv bewertet (Abbildung vgl. Abbildung 19).

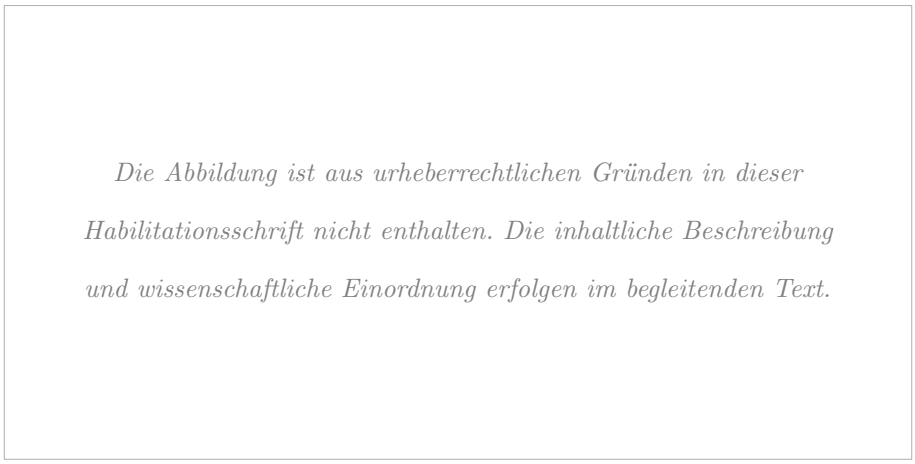

Abbildung 18: Vergleich der gemittelten Likert-skalierten Bewertungen im Vergleich zwischen Radiologen und Traumatologen (post-Intervention). Die Abbildung entspricht Table 3 aus Hoppe et al. [52].

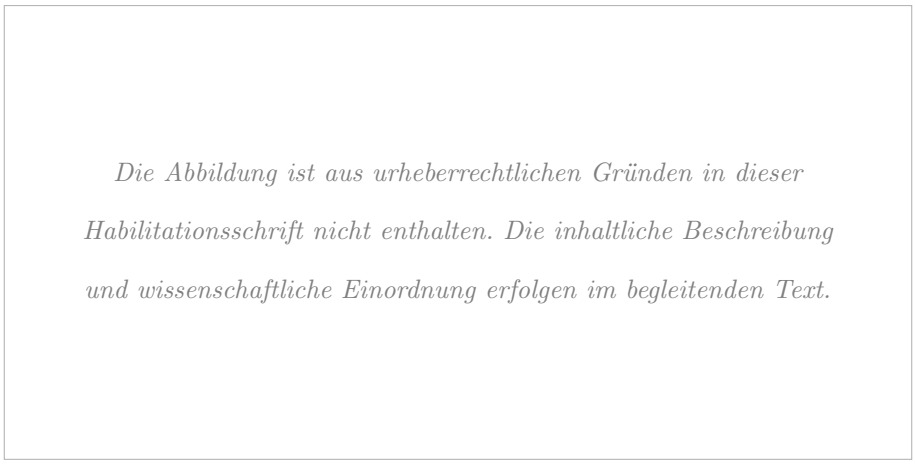

Abbildung 19: Korrelationsmatrix (Spearman-Korrelation) der Umfrage-Ergebnisse vor Intervention (oberes Dreieck der Matrix) und nach Intervention (unteres Dreieck der Matrix). Die Abbildung entspricht Figure 5 aus Hoppe et al. [52].

Die Studie zeigt, dass die Einführung von KI-Lösungen in die klinische Routine der Notfalldiagnostik eine edukative Wirkung (Umgang mit KI-Lösungen) auf die anwendenden Ärzte hat. Die Wahrnehmung von KI-Lösungen kann sich durch Einführung eines „guten Algorithmus“ (d.h. Algorithmus mit guten Ergebnissen in gut durchgeführten externen Validierungsstudien) verbessern. Gleichmaßen wird durch die klinische Anwendung das Konzept des *second readers* – d.h. dass KI primär die Befunder unterstützt statt sie zu ersetzen – unterstrichen.

3.2 Optimierung von KI-Thresholds eines Röntgen-Thorax-Algorithmus zur verbesserten KI-Wahrnehmung

Threshold optimization in AI chest radiography analysis: Integrating real-world data and clinical subgroups

Rudolph, J., Huemmer, C., Preuhs, A., Buizza, G., Dinkel, J., Koliogiannis, V., Fink, N., Goller, S.S., Schwarze, V., Heimer, M., Hoppe, B.F., Liebig, T., Ricke, J., Sabel, B.O., & Rueckel, J. (2025). Threshold optimization in AI chest radiography analysis: integrating real-world data and clinical subgroups. European radiology experimental, 9(1), 95.

https://doi.org/10.1186/s41747-025-00632-8

Journal impact factor (2024): 3.6

Diese Originalarbeit befasst sich ebenfalls mit dem Thema der KI-Wahrnehmung im klinischen Alltag. Da die Wahrnehmung und das Vertrauen in KI-Lösungen aus eigener Beobachtung und Erfahrung maßgeblich von der Performance der Algorithmen abhängt, besteht bei zunehmender klinischer Anwendung auch ein stetiger Bedarf, die Algorithmen an den persönlichen Bedarf anzupassen (z. B. Berücksichtigung besonderer Patienteneigenschaften der Klinik/Ambulanz). Eine mögliche Stellschraube ist dabei die Anpassung der optimalen Threshold: KI-Lösungen geben in der Regel pathologiespezifische *confidence scores* aus. Muss wie bei der Entscheidung über das Vorliegen einer Pathologie eine Ja-Nein-Entscheidung getroffen werden, wird eine threshold definiert, die diese dichotome Einteilung ermöglicht. In dieser Studie stellen wir eine Optimierungs-Pipeline vor, die mit der Anwendung von (1) einer gut validierten Studienkohorte (563 Röntgenbilder, sechs radiologische Referenzreader, vgl. 2.1 und 2.2) sowie (2) einer repräsentativen klinischen 1-Jahres-Kohorte (15.786 Röntgenbilder, Jahr 2018, LMU Klinikum) eine Optimierung für vier häufig auf Röntgen-Thorax-Bilder sichtbare Pathologien durchführt (Pleuraergüsse, Pneumonie-suspekte Konsolidierungen, Pneumothorax, Rundherde). Durch Subgruppenbildung in *stationäre* und *ambulante* Patienten kann dabei auch eine patientenspezifische Optimierung durchgeführt werden (vgl. Abbildung 20 - hier am Beispiel für ambulante Patienten)

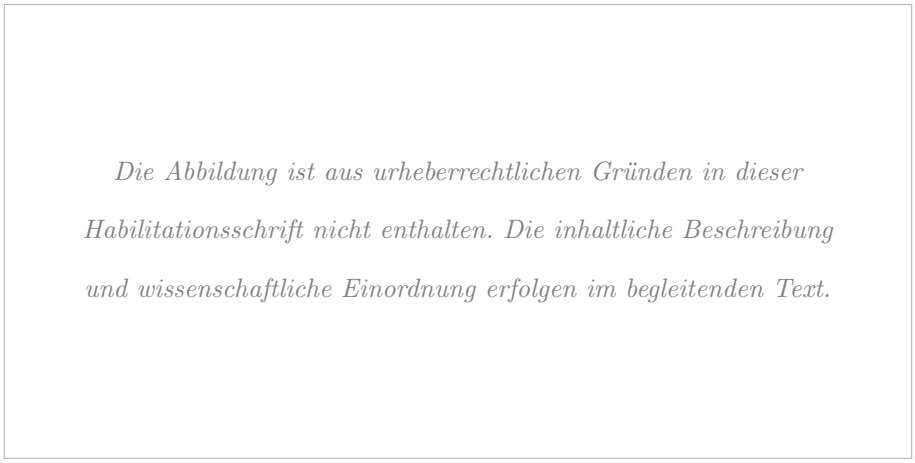

Abbildung 20: Threshold-Optimierung am Beispiel der ambulanten Patienten. Spaltenweise Darstellung verschiedener Pathologien (Pleuraerguss, Pneumonie-suspekte Konsolidierungen, Pneumothorax, Rundherde von links nach rechts). 1. Zeile: ROC-Analyse der KI-Performance in der Studiengruppe auf Basis der sechs radiologischen Referenzreader. Die dargestellten Metriken basieren auf optimierten Thresholds (OPOTs). 2. Zeile: Histogramm der KI-Scores in der ambulanten Kohorte mit markierten OPOTs der Einzelreader und dem Mittelwert (schwarz). 3. Zeile: Anteil positiver Befunde (rot) pro OPOT. 4. Zeile (Hauptinfo, dunkelrot gerahmt): Zusammenhang zwischen erreichbarer KI-Sensitivität in der Studie und resultierender *alert rate* (Anteil KI-positiver Befunde) in der ambulanten Kohorte. Die Ziel-Sensitivität ergibt sich aus dem mittleren Kurvenverlauf (schwarz), wo die Steigung erstmals >1 ist; der Zielwert und die zugehörige *alert rate* sind oben links angegeben. Die Abbildung entspricht Figure 2 aus Rudolph et al. [53].

Vergleicht man beim getesteten Algorithmus die vom KI-Hersteller angegebenen Threshold mit den von uns optimierten, so konnte das größte „Sensitivierungspotenzial“ bei der Pathologie Pleuraerguss gefunden werden. In der Gruppe der stationären Patienten ließ sich die Sensitivität von 46,8 % (bei Anwendung der Hersteller-Threshold) auf 87,2 % (bei Anwendung der optimierten Threshold, vgl. Abbildung 21) verbessern. In der Gruppe der ambulanten Patienten stieg die Sensitivität von 76,3 % auf 93,5 % (Abbildung 21).

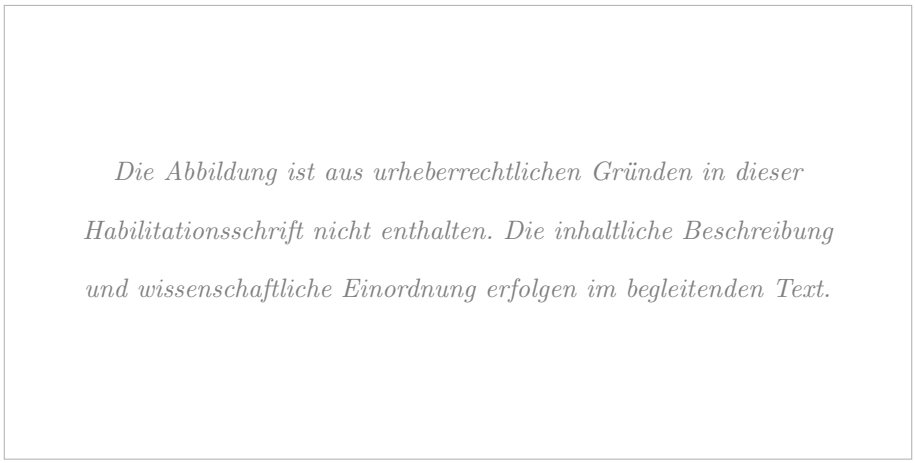

Abbildung 21: Diagnostische Metriken optimierter Thresholds (OPOT, IPOT) und der KI-Entwickler-Thresholds (AIDT) bei der Detektion von Pleuraergüssen und Pneumonie-suspekten Konsolidierungen. Die Abbildung entspricht Table 2 aus Rudolph et al. [53].

Neben weiteren Sensitivitätsverbesserungen in der Pathologie Konsolidierungen fanden wir auch umgekehrt ein „De-Sensitivierungspotenzial“ bei der Pathologie Rundherde. Hier konnte die statistische Accuracy durch Anwendung der optimierten Threshold von 69,5 % auf 82,5 % erhöht werden (stationäre Patienten, vgl. 22), während die Sensitivität keine Änderung zeigte. Bei einzelnen Subgruppen war das Optimierungspotenzial teils auch weniger ausgeprägt oder konnte aufgrund geringer statistischer Aussagekraft nur eingeschränkt quantifiziert werden (z. B. für die Pathologie Pneumothorax aufgrund der niedrigen Prävalenz in der klinischen 1-Jahres-Kohorte).

Die Abbildung ist aus urheberrechtlichen Gründen in dieser Habilitationsschrift nicht enthalten. Die inhaltliche Beschreibung und wissenschaftliche Einordnung erfolgen im begleitenden Text.

Abbildung 22: Diagnostische Metriken optimierter Thresholds (OPOT, IPOT) und der KI-Entwickler-Thresholds (AIDT) bei der Detektion von Pneumothoraces und Lungenrundherden. Analog zu Abbildung 21. Die Abbildung entspricht Table 3 aus Rudolph et al. [53].

Die vorliegende Pipeline ermöglicht eine klinische und patientenorientierte Optimierung von KI-Algorithmen und kann damit zu einer besseren Wahrnehmung der Algorithmen beitragen. Die Pipeline kann analog auch auf weitere Röntgen-Thorax-Algorithmen angewendet werden.

Abkürzungsverzeichnis

AUC Area Under The Curve

CAD Computer Aided Detection

CT Computertomographie

KI Künstliche Intelligenz

LLMs Large Language Models

MRT Magnetresonanztomographie

NLP Verabeitung natürlicher Sprache, engl. natural-language-processing

PACS Picture Archiving and Communication System

ROC Receiver Operating Characteristic

ZVKs Zentralvenöse Katheter

Literaturverzeichnis

[1] Moor, J. (2006). The Dartmouth College artificial intelligence conference: The next fifty years. *Ai Magazine*, *27*(4), 87–87.

[2] Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1980). Applications of artificial intelligence for organic chemistry: the DENDRAL project. *(McGraw-Hill Companies)*.

[3] Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1993). DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial intelligence*, *61*(2), 209–261.

[4] Oakden-Rayner, L. (2019). The Rebirth of CAD: How Is Modern AI Different from the CAD We Know? *Radiol Artif Intell*, *1*(3), e180089. https://doi.org/10.1148/ryai.2019180089

[5] Malich, A., Fischer, D. R., & Böttcher, J. (2006). CAD for mammography: the technique, results, current role and further developments. *Eur Radiol*, *16*(7), 1449–60. https://doi.org/10.1007/s00330-005-0089-x

[6] Reeves, A. P., & Kostis, W. J. (2000). COMPUTER-AIDED DIAGNOSIS FOR LUNG CANCER. *Radiologic Clinics of North America*, *38*(3), 497–509. https://doi.org/10.1016/s0033-8389(05)70180-9

[7] Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E., & Elmore, J. G. (2007). Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*, *356*(14), 1399–409. https://doi.org/10.1056/NEJMoa066099

[8] Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2015). Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern Med*, *175*(11), 1828–37. https://doi.org/10.1001/jamainternmed.2015.5231

[9] The Royal College of Radiologists. (2024). *Clinical Radiology Workforce Cencus 2024 Report* [[Online; abgerufen am 09.07.2025]]. The Royal College of Radiologists. https://www.rcr.ac.uk/media/4imb5jge/_rcr-2024-clinical-radiology-workforce-census-report.pdf

[10] Sokolovskaya, E., Shinde, T., Ruchman, R. B., Kwak, A. J., Lu, S., Shariff, Y. K., Wiggins, E. F., & Talangbayan, L. (2015). The Effect of Faster Reporting Speed for Imaging Studies on the Number of Misses and Interpretation Errors: A Pilot Study. *J Am Coll Radiol*, *12*(7), 683–8. https://doi.org/10.1016/j.jacr.2015.03.040

[11] O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. http://arxiv.org/abs/1511.08458v2;%20http://arxiv.org/pdf/1511.08458v2

[12] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

[13] Milea, D., Najjar, R. P., Zhubo, J., Ting, D., Vasseneix, C., Xu, X., Fard, M. A., Fonseca, P., Vanikieti, K., Lagrèze, W. A., Morgia, C. L., Cheung, C. Y., Hamann, S., Chiquet, C., Sanda, N., Yang, H., Mejico, L. J., Rougier, M.-B., Kho, R., … Biousse, V. (2020). Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs. *N Engl J Med*, *382*(18), 1687–1695. https://doi.org/10.1056/NEJMoa1917130

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

[15] Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., Mahajan, V., Rao, P., & Warier, P. (2018). Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet*, *392*(10162), 2388–2396. https://doi.org/10.1016/S0140-6736(18)31645-3

[16] Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*, *25*(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x

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

[18] Lotter, W., Diab, A. R., Haslam, B., Kim, J. G., Grisot, G., Wu, E., Wu, K., Onieva, J. O., Boyer, Y., Boxerman, J. L., Wang, M., Bandler, M., Vijayaraghavan, G. R., & Sorensen, A. G. (2021). Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med*, *27*(2), 244–249. https://doi.org/10.1038/s41591-020-01174-9

[19] Eisemann, N., Bunk, S., Mukama, T., Baltus, H., Elsner, S. A., Gomille, T., Hecht, G., Heywang-Köbrunner, S., Rathmann, R., Siegmann-Luz, K., Töllner, T., Vomweg, T. W., Leibig, C., & Katalinic, A. (2025). Nationwide real-world implementation of ai for cancer detection in population-based mammography screening. *Nature Medicine*, *31*(3), 917–924. https://doi.org/10.1038/s41591-024-03408-6

[20] van Leeuwen, K. G., Schalekamp, S., Rutten, M. J. C. M., van Ginneken, B., & de Rooij, M. (2021). Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol*, *31*(6), 3797–3804. https://doi.org/10.1007/s00330-021-07892-z

[21] Pauling, C., Kanber, B., Arthurs, O. J., & Shelmerdine, S. C. (2023). Commercially available artificial intelligence tools for fracture detection: The evidence. *BJR|Open*, *6*(1). https://doi.org/10.1093/bjro/tzad005

[22] Qin, Z. Z., Van der Walt, M., Moyo, S., Ismail, F., Maribe, P., Denkinger, C. M., Zaidi, S., Barrett, R., Mvusi, L., Mkhondo, N., Zuma, K., Manda, S., Koeppel, L., Mthiyane, T., & Creswell, J. (2024). Computer-aided detection of tuberculosis from chest radiographs in a tuberculosis prevalence survey in south africa: External validation and modelled impacts of commercially available artificial intelligence software. *The Lancet Digital Health*, *6*(9), e605–e613. https://doi.org/10.1016/s2589-7500(24)00118-3

[23] Salam, B., Stüwe, C., Nowak, S., Sprinkart, A. M., Theis, M., Kravchenko, D., Mesropyan, N., Dell, T., Endler, C., Pieper, C. C., Kuetting, D. L., Luetkens, J. A., & Isaak, A. (2025). Large language models for error detection in radiology reports: A comparative analysis between closed-source and privacy-compliant open-source models. *European Radiology*, *35*(8), 4549–4557. https://doi.org/10.1007/s00330-025-11438-y

[24] Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B. O., Ricke, J., & Ingrisch, M. (2023). Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *European Radiology*, *34*(5), 2817–2825. https://doi.org/10.1007/s00330-023-10213-1

[25] Bhayana, R., Alwahbi, O., Ladak, A. M., Deng, Y., Basso Dias, A., Elbanna, K., Abreu Gomez, J., Jajodia, A., Jhaveri, K., Johnson, S., Kajal, D., Wang, D., Soong, C., Kielar, A., & Krishna, S. (2025). Leveraging large language models to generate clinical histories for oncologic imaging requisitions. *Radiology*, *314*(2). https://doi.org/10.1148/radiol.242134

[26] Spitzer, P., Hendriks, D., Rudolph, J., Schlaeger, S., Ricke, J., Kühl, N., Hoppe, B. F., & Feuerriegel, S. (2025). The effect of medical explanations from large language models on diagnostic decisions in radiology. *medRxiv*. https://doi.org/10.1101/2025.03.04.25323357

[27] Wadden, J. J. (2021). Defining the undefinable: the black box problem in healthcare artificial intelligence. *J Med Ethics*. https://doi.org/10.1136/medethics-2021-107529

[28] Rueckel, J., Trappmann, L., Schachtner, B., Wesp, P., Hoppe, B. F., Fink, N., Ricke, J., Dinkel, J., Ingrisch, M., & Sabel, B. O. (2020). Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs. *Invest Radiol*, *55*(12), 792–798. https://doi.org/10.1097/RLI.0000000000000707

[29] Rueckel, J., Huemmer, C., Fieselmann, A., Ghesu, F.-C., Mansoor, A., Schachtner, B., Wesp, P., Trappmann, L., Munawwar, B., Ricke, J., Ingrisch, M., & Sabel, B. O. (2021). Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training. *Eur Radiol*, *31*(10), 7888–7900. https://doi.org/10.1007/s00330-021-07833-w

[30] Rudolph, J., Fink, N., Dinkel, J., Koliogiannis, V., Schwarze, V., Goller, S., Erber, B., Geyer, T., Hoppe, B. F., Fischer, M., Khaled, N. B., Jörgens, M., Ricke, J., Rueckel, J., & Sabel, B. O. (2021). Interpretation of Thoracic Radiography Shows Large Discrepancies Depending on the Qualification of the Physician-Quantitative Evaluation of Interobserver Agreement in a Representative Emergency Department Scenario. *Diagnostics (Basel)*, *11*(10). https://doi.org/10.3390/diagnostics11101868

[31] Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., Kahn, C. E., Abbara, S., Afat, S., Anazodo, U. C., Andreychenko, A., Asselbergs, F. W., Badano, A., Baessler, B., Bold, B., Bisdas, S., Brismar, T. B., Cacciamani, G. E., Carrino, J. A., ... Zins, M. (2024). Checklist for artificial intelligence in medical imaging (claim): 2024 update. *Radiology: Artificial Intelligence*, *6*(4). https://doi.org/10.1148/ryai.240300

[32] Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., & Rosenow, E. C. (2012). Interpretation of plain chest roentgenogram. *Chest*, *141*(2), 545–558. https://doi.org/10.1378/chest.10-1302

[33] Gurney, J. W. (1995). Why chest radiography became routine. *Radiology*, *195*(1), 245–6. https://doi.org/10.1148/radiology.195.1.7892479

[34] Rudolph, J., Huemmer, C., Ghesu, F.-C., Mansoor, A., Preuhs, A., Fieselmann, A., Fink, N., Dinkel, J., Koliogiannis, V., Schwarze, V., Goller, S., Fischer, M., Jörgens, M., Khaled, N. B., Vishwanath, R. S., Balachandran, A., Ingrisch, M., Ricke, J., Sabel, B. O., & Rueckel, J. (2022). Artificial Intelligence in

Chest Radiography Reporting Accuracy: Added Clinical Value in the Emergency Unit Setting Without 24/7 Radiology Coverage. *Invest Radiol*, *57*(2), 90–98. https://doi.org/10.1097/RLI.0000000000000813

[35] Rudolph, J., Huemmer, C., Preuhs, A., Buizza, G., Hoppe, B. F., Dinkel, J., Koliogiannis, V., Fink, N., Goller, S. S., Schwarze, V., Mansour, N., Schmidt, V. F., Fischer, M., Jörgens, M., Ben Khaled, N., Liebig, T., Ricke, J., Rueckel, J., & Sabel, B. O. (2024). Nonradiology health care professionals significantly benefit from ai assistance in emergency-related chest radiography interpretation. *CHEST*, *166*(1), 157–170. https://doi.org/10.1016/j.chest.2024.01.039

[36] Kunz, W. G., Patzig, M., Crispin, A., Stahl, R., Reiser, M. F., & Notohamiprodjo, M. (2018). The Value of Supine Chest X-Ray in the Diagnosis of Pneumonia in the Basal Lung Zones. *Acad Radiol*, *25*(10), 1252–1256. https://doi.org/10.1016/j.acra.2018.01.027

[37] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. http://arxiv.org/abs/1711.05225v3;%20http://arxiv.org/pdf/1711.05225v3

[38] Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Patel, B. N., Yeom, K. W., Shpanskaya, K., Blankenberg, F. G., Seekins, J., Amrhein, T. J., Mong, D. A., Halabi, S. S., Zucker, E. J., ... Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*, *15*(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686

[39] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2017.369

[40] Rudolph, J., Schachtner, B., Fink, N., Koliogiannis, V., Schwarze, V., Goller, S., Trappmann, L., Hoppe, B. F., Mansour, N., Fischer, M., Khaled, N. B., Jörgens, M., Dinkel, J., Kunz, W. G., Ricke, J., Ingrisch, M., Sabel, B. O., & Rueckel, J. (2022). Clinically focused multi-cohort benchmarking as a tool for external validation of artificial intelligence algorithm performance in basic chest radiography analysis. *Sci Rep*, *12*(1), 12764. https://doi.org/10.1038/s41598-022-16514-7

[41] McGee, D. C., & Gould, M. K. (2003). Preventing complications of central venous catheterization. *N Engl J Med*, *348*(12), 1123–33. https://doi.org/10.1056/NEJMra011883

[42] Lockwood, J., & Desai, N. (2019). Central venous access. *Br J Hosp Med (Lond)*, *80*(8), C114–C119. https://doi.org/10.12968/hmed.2019.80.8.C114

[43] Stone, P. A., Hass, S. M., Knackstedt, K. S., & Jagannath, P. (2012). Malposition of a central venous catheter into the right internal mammary vein: review of complications of catheter misplacement. *Vasc Endovascular Surg*, *46*(2), 187–9. https://doi.org/10.1177/1538574411433288

[44] Rueckel, J., Huemmer, C., Shahidi, C., Buizza, G., Hoppe, B. F., Liebig, T., Ricke, J., Rudolph, J., & Sabel, B. O. (2024). Artificial Intelligence to Assess Tracheal Tubes and Central Venous Catheters in Chest Radiographs Using an Algorithmic Approach With Adjustable Positioning Definitions. *Invest Radiol*, *59*(4), 306–313. https://doi.org/10.1097/RLI.0000000000001018

[45] Friedlinger, M., Schad, L. R., Blüml, S., Tritsch, B., & Lorenz, W. J. (1995). Rapid automatic brain volumetry on the basis of multispectral 3D MR imaging data on personal computers. *Comput Med Imaging Graph*, *19*(2), 185–205. https://doi.org/10.1016/0895-6111(94)00045-e

[46] Kitagaki, H., Mori, E., Yamaji, S., Ishii, K., Hirono, N., Kobashi, S., & Hata, Y. (1998). Frontotemporal dementia and Alzheimer disease: evaluation of cortical atrophy with automated hemispheric surface display generated with MR images. *Radiology*, *208*(2), 431–9. https://doi.org/10.1148/radiology.208.2.9680572

[47] Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*(2), 774–81. https://doi.org/10.1016/j.neuroimage.2012.01.021

[48] Sargolzaei, S., Sargolzaei, A., Cabrerizo, M., Chen, G., Goryawala, M., Pinzon-Ardila, A., Gonzalez-Arias, S. M., & Adjouadi, M. (2015). Estimating Intracranial Volume in Brain Research: An Evaluation of Methods. *Neuroinformatics*, *13*(4), 427–41. https://doi.org/10.1007/s12021-015-9266-5

[49] Rudolph, J., Rueckel, J., Döpfert, J., Ling, W. X., Opalka, J., Brem, C., Hesse, N., Ingenerf, M., Koliogiannis, V., Solyanik, O., Hoppe, B. F., Zimmermann, H., Flatz, W., Forbrig, R., Patzig, M., Rauchmann, B.-S., Perneczky, R., Peters, O., Priller, J., … Stoecklein, S. (2024). Artificial intelligence–based rapid brain volumetry substantially improves differential diagnosis in dementia. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *16*(4). https://doi.org/10.1002/dad2.70037

[50] Duron, L., Ducarouge, A., Gillibert, A., Lainé, J., Allouche, C., Cherel, N., Zhang, Z., Nitche, N., Lacave, E., Pourchot, A., Felter, A., Lassalle, L., Regnard, N.-E., & Feydy, A. (2021). Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. *Radiology*, *300*(1), 120–129. https://doi.org/10.1148/radiol.2021203886

[51] Guermazi, A., Tannoury, C., Kompel, A. J., Murakami, A. M., Ducarouge, A., Gillibert, A., Li, X., Tournier, A., Lahoud, Y., Jarraya, M., Lacave, E., Rahimi, H., Pourchot, A., Parisien, R. L., Merritt, A. C., Comeau, D., Regnard, N.-E., & Hayashi, D. (2022). Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. *Radiology*, *302*(3), 627–636. https://doi.org/10.1148/radiol.210937

[52] Hoppe, B. F., Rueckel, J., Dikhtyar, Y., Heimer, M., Fink, N., Sabel, B. O., Ricke, J., Rudolph, J., & Cyran, C. C. (2024). Implementing Artificial Intelligence for Emergency Radiology Impacts Physicians' Knowledge and Perception: A Prospective Pre- and Post-Analysis. *Invest Radiol*, *59*(5), 404–412. https://doi.org/10.1097/RLI.0000000000001034

[53] Rudolph, J., Huemmer, C., Preuhs, A., Buizza, G., Dinkel, J., Koliogiannis, V., Fink, N., Goller, S. S., Schwarze, V., Heimer, M., Hoppe, B. F., Liebig, T., Ricke, J., Sabel, B. O., & Rueckel, J. (2025). Threshold optimization in ai chest radiography analysis: Integrating real-world data and clinical subgroups. *European Radiology Experimental*, *9*(1). https://doi.org/10.1186/s41747-025-00632-8

Danksagung

- Mein herzlicher Dank gilt **Prof. Dr. Jens Ricke** für die Möglichkeit, meine Forschung und Habilitation in der Klinik und Poliklinik für Radiologie an der LMU München durchzuführen, für die Bereitstellung der notwendigen Mittel und Ressourcen sowie für die Übernahme der Rolle als Vorsitzender des Fachmentorats.

- Weiterhin danke ich **Prof. Dr. Bastian Sabel** für die Betreuung zahlreicher wissenschaftlicher Arbeiten und seine wertvolle Mitwirkung im Fachmentorat, **Prof. Dr. Christian Stief** als weiterem Mitglied des Fachmentorats sowie **Prof. Dr. Diane Renz** und **Prof. Dr. Thorsten Bley** für die externe Begutachtung.

- Ein besonderer Dank gilt meinen wissenschaftlichen Kollaborationspartnern, insbesondere **PD Dr. Johannes Rückel** und **Dr. Boj Hoppe**, mit denen ein kontinuierlicher, tiefgehender wissenschaftlicher Austausch stattgefunden hat und die an laufenden Projekten beteiligt sind.

- Ebenso danke ich allen weiteren Kollaborationspartnern, darunter **Dr. Balthasar Schachtner**, **Prof. Dr. Michael Ingrisch**, **Prof. Dr. Clemens Cyran**, **Prof. Dr. Sophia Stöcklein**, den Partnern am **DZNE** sowie den industriellen Partnern von **Siemens Healthineers, mediaire** und **deepc**.

- Mein Dank gilt auch **allen Mitarbeitern der Klinik und Poliklinik für Radiologie**, die an der Erhebung von Bild- und Auswertungsmaterial mitgewirkt haben, insbesondere den MTRs und wissenschaftlichen Mitarbeitern.

- Ebenso danke ich **allen hier nicht namentlich erwähnten Kollaborationspartnern**, die an den erwähnten sowie an aktuell laufenden (nicht hier dargestellten) Studien beteiligt sind.

- Meinem ehemaligen Chef **Prof. Dr. Thomas Helmberger** danke ich sehr dafür, dass er mich dabei unterstützt hat, den Schritt in die universitäre Radiologie zu wagen, und mir dabei als wichtiger Mentor zur Seite stand.

- Ein besonderer Dank gilt **dem gesamten Team der Kinderradiologie an der LMU**, wo ich zuletzt meine Spezialisierung in Kinder- und Jugendradiologie durchgeführt habe und das mich sowohl fachlich als auch menschlich sehr stark unterstützt hat.

- Nicht zuletzt gilt mein besonderer Dank meinen Freunden und meiner Familie, die mich in allen Lebenslagen unterstützt haben – insbesondere meiner Partnerin **Esa** und unserem Sohn **Jonathan**, denen ich diese Arbeit widme.