Aus der Klinik und Poliklinik für Radiologie
Klinikum der Ludwig-Maximilians-Universität München
Direktor: Prof. Dr. med. Jens Ricke

Application of Machine Learning in CT Colonography and Radiological Age Assessment: Enhancing Traditional Diagnostics in Radiology

Kumulative Dissertation
zum Erwerb des Doktorgrades der Naturwissenschaften (Dr. rer. nat.)
an der Medizinischen Fakultät der
Ludwig–Maximilians–Universität zu München

vorgelegt von
Philipp Wesp
aus
Mannheim

2023

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Betreuer:	Prof. Dr. rer. nat. Michael Ingrisch
Zweitgutachter:	Prof. Dr. rer. nat. Guido Böning
Dekan:	Prof. Dr. med. Thomas Gudermann
Tag der mündlichen Prüfung:	04. März 2024

Eidesstattliche Versicherung

Wesp, Philipp

Name, Vorname

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Titel

Application of Machine Learning in CT Colonography and Radiological Age Assessment: Enhancing Traditional Diagnostics in Radiology

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht habe und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, den 13. März 2024 — Philipp Wesp

Ort, Datum — Unterschrift

Contents

List of Abbreviations

2D	Two dimensional
3D	Three dimensional
ADAM	Adaptive moment estimation
AI	Artificial intelligence
AGFAD	Study Group on Forensic Age Diagnostics
AP	Average precision
AUC	Area under the curve
CNN	Convolutional neural network
CRC	United Nations Convention on the Rights of the Child
CT	Computed Tomography
CTC	Computed Tomography colonography
DL	Deep learning
EID	Energy-integrating detector
EUAA	European Union Agency for Asylum
FC	Fully connected
GLCM	Gray-level co-occurrence matrix
GLRLM	Gray-level run-length matrix
HU	Houndsfield units
IBSI	Image Biomarker Standardization Initiative
IoU	Intersection over Union
LoG	Laplacian of Gaussian
MAE	Mean absolute error
MDI	Mean decrease in impurity
ML	Machine learning
MRI	Magnetic resonance imaging
PACS	Picture archiving and communication systems
PCD	Photon-counting detector
PET	Positron emission tomography
ROC	Receiver operating characteristic
ReLU	Rectified linear unit
RIS	Radiology information systems
ROI	Region of interest

SD	Standard deviation
SGD	Stochastic gradient descent
SOI	Structure of interest
TCIA	The Cancer Imaging Archive

List of Publications

Original Publications

The following three original publications are subject of this cumulative dissertation in accordance with the promotion regulation for natural sciences of the medical faculty of Ludwig-Maximilians-University Munich.

[1] S. Grosu, **P. Wesp**, A. Graser, S. Maurus, C. Schulz, T. Knösel, C. C. Cyran, J. Ricke, M. Ingrisch, and P. M. Kazmierczak, "Machine Learning-based Differentiation of Benign and Premalignant Colorectal Polyps Detected with CT Colonography in an Asymptomatic Screening Population: A Proof-of-Concept Study," *Radiology*, p. 202 363, Feb. 2021, ISSN: 0033-8419. DOI: 10.1148/radiol.2021202363.

[2] **P. Wesp**, S. Grosu, A. Graser, S. Maurus, C. Schulz, T. Knösel, M. P. Fabritius, B. Schachtner, B. M. Yeh, C. C. Cyran, J. Ricke, P. M. Kazmierczak, and M. Ingrisch, "Deep learning in CT colonography: Differentiating premalignant from benign colorectal polyps," *European Radiology*, vol. 32, no. 7, pp. 4749–4759, Jul. 2022, ISSN: 1432-1084. DOI: 10.1007/s00330-021-08532-2.

[3] **P. Wesp**, B. O. Sabel, A. Mittermeier, A. T. Stüber, K. Jeblick, P. Schinke, M. Mühlmann, F. Fischer, R. Penning, J. Ricke, M. Ingrisch, and B. M. Schachtner, "Automated localization of the medial clavicular epiphyseal cartilages using an object detection network: A step towards deep learning-based forensic age assessment," *International Journal of Legal Medicine*, Feb. 2023. DOI: 10.1007/s00414-023-02958-7.

Complementing Publication Accepted After Submission

The following original publication complements this cumulative dissertation. The article was under review in a peer-reviewed academic journal at the time of submission and was published prior to the day of the oral dissertation examination.

[1] **P. Wesp**, B. M. Schachtner, K. Jeblick, J. Topalis, M. Weber, F. Fischer, R. Penning, J. Ricke, M. Ingrisch, and B. O. Sabel, "Radiological age assessment

based on clavicle ossification in ct: Enhanced accuracy through deep learning," *International Journal of Legal Medicine*, Jan. 2024. DOI: 10.1007/s00414-024-03167-6.

Additional Publications

I contributed to the following publications as a co-author while pursuing my dissertation at the medical faculty of Ludwig-Maximilians-University Munich.

[1] J. Rueckel, L. Trappmann, B. Schachtner, **P. Wesp**, B. F. Hoppe, N. Fink, J. Ricke, J. Dinkel, M. Ingrisch, and B. O. Sabel, "Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs," *Investigative Radiology*, vol. 55, no. 12, pp. 792–798, Dec. 2020. DOI: 10.1097/RLI.0000000000000707.

[2] J. Rueckel, C. Huemmer, A. Fieselmann, F.-c. Ghesu, A. Mansoor, B. Schachtner, **P. Wesp**, L. Trappmann, B. Munawwar, J. Ricke, and M. Ingrisch, "Pneumothorax detection in chest radiographs: Optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training," *European Radiology*, vol. 31, no. 10, pp. 7888–7900, Oct. 2021. DOI: 10.1007/s00330-021-07833-w.

[3] A. Mittermeier, P. Reidler, M. P. Fabritius, B. Schachtner, **P. Wesp**, B. Ertl-Wagner, O. Dietrich, J. Ricke, L. Kellert, S. Tiedt, W. G. Kunz, and M. Ingrisch, "End-to-End Deep Learning Approach for Perfusion Data: A Proof-of-Concept Study to Classify Core Volume in Stroke CT," *Diagnostics*, vol. 12, no. 5, May 2022. DOI: 10.3390/diagnostics12051142.

[4] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, **P. Wesp**, B. Sabel, J. Ricke, and M. Ingrisch, *ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports*, Dec. 2022. DOI: 10.48550/arXiv.2212.14882.

Conference Proceedings, Presentations and Posters

I contributed with the following abstracts, presentations and/or posters to scientific conferences while pursuing my dissertation at the medical faculty of Ludwig-Maximilians-University Munich.

[1] **P. Wesp**, M. Ingrisch, A. Graser, S. Maurus, C. Schulz, T. Knoesel, C. Cyran, J. Ricke, P. Kazmierczak, and S. Grosu, "Machine Learning-basierte Klassifizierung von Kolonläsionen in der CT-Kolonografie: Ein Radiomics-Ansatz," ser. RöFo 2020, vol. 192, 2020, pp. 79–79. DOI: 10.1055/s-0040-1703343.

[2] **P. Wesp**, L. Libon, L. Voland, J. van Voorden, J. Wu, C. Cyran, B. Schachtner, M. Ingrisch, and B. Sabel, "CNN-basierte forensische Altersschätzung anhand von CT-Untersuchungen der Epiphysenfugen der Claviculae: Ein Schüler-

beitrag zum BWKI Finale 2020," ser. RöFo 2021, vol. 193, 2021, pp. 4–5. DOI: 10.1055/s-0041-1723138.

[3] **P. Wesp**, S. Grosu, A. Mittermeier, A. T. Stüber, S. Maurus, J. Rückel, C. C. Cyran, B. Sabel, J. Ricke, B. Schachtner, and M. Ingrisch, "Improving automated clinical decision making by allowing machine-learning models to say "I don't know": Balancing performance against abstention [presentation]," in *RSNA 2021*, Chicago, IL, USA: RSNA, Nov. 2021.

[4] **P. Wesp**, B. M. Schachtner, S. Grosu, A. Mittermeier, A. T. Stüber, C. C. Cyran, and J. Ricke, "Allowing machine learning models to say "I don't know": Improving automated clinical decision-making by balancing performance against abstention," in *ECR 2022*, Vienna, Austria: European Society of Radiology, Vienna, Austria, Jul. 2022. DOI: 10.26044/ecr2022/C-14829.

Abstract

Machine learning has the potential to overcome challenges in radiology where traditional diagnostic methods reach their limits. This work addresses two such challenging clinical problems from two areas of radiology using different machine learning approaches. First, differentiating premalignant from benign colorectal polyps in computed tomography (CT) colonography. Second, continuous age predictions for radiological age assessment based on clavicle ossification in CT.

The first clinical problem regards the differentiation of colorectal polyps to prevent colorectal cancer, which is among the three leading causes of cancer-related death in industrialized countries. CT colonography is a non-invasive screening method for colorectal cancer that can reliably detect polyps. However, it cannot distinctively differentiate benign polyps from premalignant ones that can turn into cancer. This work aims to enable this differentiation of colorectal polyps using machine learning.

A training dataset was acquired in a secondary analysis of a previous prospective trial. First, colorectal polyps of all size categories and morphologies were manually segmented in CT colonography scans and polyps were classified as benign (hyperplastic polyp or regular mucosa) or premalignant (adenoma) according to the histopathologic reference standard. The assembled training dataset consisted of 107 colorectal polyps in 63 patients and 169 manual polyp segmentation masks in CT colonography scans. Next, radiomic image features characterizing shape (n = 14), gray level histogram statistics (n = 18), and image texture (n = 68) were calculated from the segmented polyps after applying 22 image filters, resulting in 1906 feature-filter combinations. Based on these features, a random forest classification model was trained on the training set to predict the polyp character. Model performance was validated in an external test dataset from a large North American multicenter CT colonography screening trial that has been made publicly accessible via The Cancer Imaging Archive. The test dataset consisted of 77 polyps in 59 patients and 118 manual polyp segmentation masks.

Random forest predictions for polyp class in the external test dataset had an area under the receiver operating characteristic curve (ROC-AUC) of 0.91, 82 % sensitivity, and 85 % specificity. These results demonstrate that machine learning enables the

non-invasive differentiation of benign and premalignant colorectal polyps with CT colonography. Consequently, this allows for individual risk stratification and therapy guidance through a more precise selection of patients who would benefit from endoscopic polypectomy.

However, the radiomics approach is impracticable for integration into everyday clinical workflows, because the manual polyp segmentation is time-consuming, expensive, and has high inter-reader variability. Therefore, two convolutional neuronal network (CNN) ensembles, SEG and noSEG, were trained on 3D CT colonography image subvolumes from the same training set to predict the polyp class. Model SEG was additionally trained with polyp segmentation masks. Diagnostic performance was validated in the same external multicentre test dataset. Additionally, predictions were analyzed with the gradient-based CNN visualization technique Grad-CAM++.

Model SEG achieved a ROC-AUC of 0.83 and 80 % sensitivity at 69 % specificity for differentiating premalignant from benign polyps. Model noSEG yielded a ROC-AUC of 0.75, 80 % sensitivity at 44 % specificity, and an average Grad-CAM++ heatmap score of $\geq$ 0.25 in 90 % of polyp tissue. These results show that deep learning also enables differentiating premalignant from benign colorectal polyps found in CT colonography scans when no segmentation mask is provided. The deep learning model noSEG learned to focus on polyp tissue for predictions without the need for prior polyp segmentation by experts. Thus, deep learning provides the basis for a fully automated CT colonography evaluation, as CNN polyp classification could be combined with already established computer-aided detection algorithms for polyp detection.

The second clinical problem regards radiological age assessment, a method for assessing a person's chronological age when the age is unknown or in serious doubt. One particular assessment approach is to examine the ossification status of the medial clavicular epiphyseal cartilages in dedicated CT scans. Next, the ossification is compared to the skeletal maturation of case groups from a reference study with known age. The inherent problem with that approach is the limited number of ossification stages that can be assessed with the human eye, which leads to a small set of discrete age estimates that can be assigned to a person. Consequently, the accuracy of these estimates is limited. To address this issue, this work investigates enabling continuous age prediction through a deep learning model that maps a thoracic CT scan to chronological age.

Training a deep learning model to solve this task on a full CT scan is challenging and requires extremely large datasets and computing resources. To lower the complexity and reduce the required resources, the first goal was to crop thoracic CT scans around the relevant structure of interest (SOI), the sternoclavicular joints. This SOI serves as an easy-to-identify proxy for the medial clavicular epiphyseal cartilages. To this end, an instance of the object detection network RetinaNet was trained to automatically locate the SOI in CT scans. This is crucial as manual SOI localization

by experts would pose a bottleneck for creating the necessary large dataset required to train the deep learning model, even when cropped around the relevant structures. Therefore, CT slices containing the SOI were manually annotated with bounding boxes around the SOI. The training dataset contained 29,656 slices from 100 CT scans of 82 different patients. The test dataset included 30,846 slices from 110 CT scans of 110 different patients. All slices in the training set were used to train the RetinaNet. Afterwards, the network was applied individually to all slices of the test dataset for SOI detection. The bounding box and slice position of the detection with the highest classification score was used as the location estimate for the SOI inside the CT scan.

The deep learning-based location estimate for the SOI was in a correct slice in 97/110 (88 %), misplaced by one slice in 5/110 (5 %), and not available in 8/110 (7 %) test scans. Also, no location estimate was misplaced by more than one slice. These results demonstrate an automated approach for annotating the medial clavicular epiphyseal cartilages, which allows creating large training and test datasets for the development of a deep learning model for radiological age assessment.

Building on the automated detection approach, a deep learning model for radiological age assessment was developed. Therefore, thoracic CT scans were retrospectively collected from the LMU University Hospital's picture archiving and communication system. Individuals aged 15.0 to 30.0 years examined in routine clinical practice were included. All scans were automatically cropped around the medial clavicular epiphyseal cartilages using the previously trained RetinaNet. The trainig dataset contained 4,400 scans of 1,935 patients and the test dataset 300 scans of 300 patients with a balanced age and sex distribution. An adaption of the popular neural network ResNet was trained to predict a person's chronological age based on these scans. In order to evaluate model performance, this work introduces an optimistic human reader performance estimate for an established reference study method for radiological age assessment.

The mean absolute error (MAE) of deep learning model predictions for chronological age was 1.65 years, and the highest observed absolute error was 6.40 years for females and 7.32 years for males. However, performance in these high-error cases could be attributed to norm-variants or pathologic disorders. The mean absolute error (MAE) of the human reader estimate was 1.84 years and the highest calculated absolute error was 3.40 years for females and 3.78 years for males. These results demonstrate that the developed deep learning approach for continuous age prediction on CT volumes showing the clavicles outperforms the human reader estimate on average.

In summary, this work demonstrates proof-of-concept machine learning approaches that address two clinical problems in radiology: colorectal cancer screening with CT colonography and radiological age assessment based on clavicle ossification in CT. The approaches successfully solved challenging problems that are otherwise difficult to overcome for conventional imaging diagnostics.

Zusammenfassung

Maschinelles Lernen hat das Potenzial Herausforderungen in der Radiologie zu bewältigen bei denen herkömmliche Diagnosemethoden an ihre Grenzen stoßen. Diese Arbeit behandelt zwei anspruchsvolle klinische Probleme aus zwei Bereichen der Radiologie unter Verwendung verschiedener Ansätze des maschinellen Lernens. Erstens, die Unterscheidung zwischen prämalignen und benignen kolorektalen Polypen in der Computertomographie (CT) Kolonographie. Zweitens, kontinuierliche Altersvorhersagen für die radiologische Altersbestimmung auf Grundlage der Verknöcherung des Schlüsselbeins in der CT.

Das erste klinische Problem ist die Unterscheidung kolorektaler Polypen im Rahmen der Darmkrebsvorsorge. Darmkrebs zählt in Industrieländern zu den drei häufigsten krebsbedingten Todesursachen. Die CT-Kolonographie ist eine nicht-invasive Methode zur Früherkennung von Darmkrebs, mit der Polypen zuverlässig erkannt werden können. Damit lässt sich jedoch nicht eindeutig zwischen gutartigen Polypen und prämalignen Polypen, welche sich zu Krebs entwickeln können, unterscheiden. Diese Arbeit hat das Ziel diese Unterscheidung von kolorektalen Polypen durch maschinelles Lernen zu ermöglichen.

Ein Trainingsdatensatz wurde im Rahmen einer Sekundäranalyse einer früheren prospektiven Studie angefertigt. Zunächst wurden kolorektale Polypen aller Größenkategorien und Morphologien in CT-Kolonographie-Scans manuell segmentiert und die Polypen gemäß dem histopathologischen Referenzstandard als gutartig (hyperplastischer Polyp oder normale Mukosa) oder prämaligne (Adenom) klassifiziert. Der Trainingsdatensatz bestand aus 107 kolorektalen Polypen von 63 Patienten und 169 manuellen Polypen-Segmentierungsmasken in CT-Kolonographie-Scans. Aus den segmentierten Polypen wurden nach Anwendung von 22 Bildfiltern radiologische Bildmerkmale berechnet, die Form (n = 14), Graustufenhistogramm-Statistik (n = 18) und Bildtextur (n = 68) charakterisieren, was insgesamt 1906 Merkmals-Filter-Kombinationen ergab. Auf Grundlage dieser Merkmale wurde ein Random-Forest-Klassifizierungsmodell auf dem Trainingssatz trainiert, um den Polypencharakter vorherzusagen. Die Unterscheidungsfähigkeit des Modells wurde anhand eines externen Testdatensatzes aus einer großen nordamerikanischen multizentrischen CT-Kolonographie-Screeningstudie validiert, die über das Cancer Imaging Archive öf-

fentlich zugänglich gemacht wurde. Der Testdatensatz bestand aus 77 Polypen von 59 Patienten und 118 manuellen Polypensegmentierungsmasken.

Die Random-Forest-Vorhersagen für die Polypenklasse im externen Testdatensatz hatten eine Fläche unter der Receiver Operating Characteristic Curve (ROC-AUC) von 0,91, eine Sensitivität von 82 % und eine Spezifität von 85 %. Diese Ergebnisse zeigen, dass maschinelles Lernen die nicht-invasive Differenzierung von benignen und prämalignen kolorektalen Polypen mit der CT-Kolonographie ermöglicht. Eine genauere Auswahl von Patienten die von einer endoskopischen Polypektomie profitieren würden, ermöglicht eine individuelle Risikostratifizierung und Therapieführung.

Der Radiomics-Ansatz ist für die Integration in den klinischen Alltag jedoch nicht praktikabel, da die manuelle Polypensegmentierung zeitaufwändig und teuer ist und eine hohe Variabilität zwischen den radiologischen Leserinnen und Lesern aufweist. Daher wurden zwei Ensembles von Convolutional Neural Networks (CNN), SEG und noSEG, auf 3D-CT-Kolonographie-Subvolumina aus demselben Trainingssatz trainiert, um die Polypenklasse vorherzusagen. Das Modell SEG wurde zusätzlich mit Polypen-Segmentierungsmasken trainiert. Die Fähigkeit korrekte Diagnosen zu erstellen wurde mit demselben externen multizentrischen Testdatensatz validiert. Zusätzlich wurden die Vorhersagen mit der gradientenbasierten CNN-Visualisierungstechnik Grad-CAM++ analysiert.

Das Modell SEG erreichte eine ROC-AUC von 0,83 und 80 % Sensitivität bei 69 % Spezifität für die Unterscheidung zwischen prämalignen und benignen Polypen. Das Modell noSEG lieferte eine ROC-AUC von 0,75, 80 % Sensitivität bei 44 % Spezifität und einen durchschnittlichen Grad-CAM++ Heatmap-Wert von $\geq 0{,}25$ bei 90 % des Polypengewebes. Diese Ergebnisse zeigen, dass Deep Learning auch dann eine Unterscheidung zwischen prämalignen und benignen kolorektalen Polypen ermöglicht, wenn keine Segmentierungsmaske vorhanden ist. Das Deep-Learning-Modell noSEG hat gelernt, sich für Vorhersagen auf Polypengewebe zu konzentrieren, ohne dass eine vorherige Segmentierung der Polypen durch Experten erforderlich ist. Deep Learning bietet somit die Grundlage für eine vollautomatische CT-Kolonographie-Auswertung, da die CNN-Polypenklassifizierung mit bereits etablierten computergestützten Algorithmen zur Polypenerkennung kombiniert werden könnte.

Das zweite klinische Problem betrifft die radiologische Altersbestimmung, eine Methode zur Schätzung des chronologischen Alters einer Person, wenn das Alter unbekannt ist oder ernsthaft angezweifelt wird. Eine bestimmte Schätzungsmethode ist die Untersuchung des Verknöcherungsstatus der Epiphysenknorpel des mittleren Schlüsselbeins in CT-Scans. Die Verknöcherung wird kategorisiert und anschließend mit der Skelettreifung von Fallgruppen aus einer Referenzstudie mit bekanntem Alter verglichen. Das inhärente Problem bei diesem Ansatz ist die begrenzte Anzahl von Verknöcherungsstadien, die mit dem menschlichen Auge beurteilt werden können, was zu einer kleinen Anzahl von diskreten Altersschätzungen führt, die einer Person zugeordnet werden können. Folglich ist die Genauigkeit dieser Schätzungen

begrenzt. Daher wird in dieser Arbeit die Möglichkeit einer kontinuierlichen Altersvorhersage durch ein Deep-Learning-Modell untersucht, das einen Thorax-CT-Scan auf das chronologische Alter abbildet.

Das Training eines Deep-Learning-Modells zur kontinuierlichen Altersvorhersage auf einem vollständigen CT-Scan ist eine Herausforderung und erfordert extrem große Datensätze und Rechenressourcen. Um die Komplexität der Problemstellung zu verringern und die erforderlichen Ressourcen zu reduzieren, bestand das erste Ziel darin, Thorax-CT-Scans um die relevante Struktur von Interesse (SOI), die Sternoklavikulargelenke, herum auszuschneiden. Diese SOI dient als einfach zu identifizierende Stellvertreterregion für die Epiphysenknorpel des medialen Schlüsselbeins. Zu diesem Zweck wurde eine Instanz des Objekterkennungsnetzes RetinaNet darauf trainiert, die SOI in CT-Scans automatisch zu lokalisieren. Dieser Schritt ist von entscheidender Bedeutung, da die manuelle Lokalisierung der SOI durch Experten einen Engpass für die Erstellung des erforderlichen großen Datensatzes darstellen würde, der für das Training eines Deep-Learning-Modells benötigt wird, selbst wenn die relevanten Strukturen ausgeschnitten werden. Daher wurden die CT-Schichten, die die SOI enthielten, manuell mit quadratischen Kästchen um die SOI herum markiert. Der Trainingsdatensatz enthielt 29.656 Schichten aus 100 CT-Scans von 82 verschiedenen Patienten. Der Testdatensatz umfasste 30.846 Schichten von 110 CT-Scans von 110 verschiedenen Patienten. Alle Schichten des Trainingsdatensatzes wurden für das Training des RetinaNet verwendet. Anschließend wurde das Netzwerk einzeln auf allen Schichten des Testdatensatzes zur SOI-Erkennung angewendet. Das Kästchen und die Schichtposition der Erkennung mit der höchsten Klassifizierungspunktzahl wurden als Schätzung für die Position der SOI innerhalb des CT-Scans verwendet.

Die auf Deep Learning basierende Positionsschätzung für die SOI befand sich in 97/110 (88 %) in einer korrekten Schicht, war in 5/110 (5 %) eine Schicht daneben und in 8/110 (7 %) Testscans nicht verfügbar. Außerdem war keine Positionsschätzung um mehr als eine Schicht verschoben. Diese Ergebnisse zeigen einen automatisierten Ansatz für die Lokalisierung der medialen klavikulären Epiphysenknorpel, welcher die Erstellung großer Trainings- und Testdatensätze für die Entwicklung eines Deep-Learning-Modells zur radiologischen Altersbestimmung ermöglicht.

Aufbauend auf dem automatischen Erkennungsansatz wurde ein Deep-Learning-Modell für die radiologische Altersbestimmung entwickelt. Dazu wurden Thorax-CT-Aufnahmen retrospektiv aus dem Bildarchivierungs- und Kommunikationssystem des Universitätsklinikums der LMU gesammelt. Eingeschlossen wurden Personen im Alter von 15,0 bis 30,0 Jahren, die in der klinischen Routinepraxis untersucht wurden. Alle Scans wurden mit Hilfe des zuvor trainierten RetinaNet automatisch um die medialen Epiphysenknorpel des Schlüsselbeins augeschnitten. Der Trainingsdatensatz enthielt 4.400 Scans von 1.935 Patienten und der Testdatensatz 300 Scans von 300 Patienten mit einer ausgewogenen Alters- und Geschlechtsverteilung. Eine an-

gepasste Version des bekannten neuronalen Netzes ResNet wurde trainiert, um das chronologische Alter einer Person auf der Grundlage dieser Scans vorherzusagen. Um die Genaugikeit des Modells besser bewerten zu können, wird in dieser Arbeit eine optimistische Schätzung der Genauigkeit einer etablierte Referenzstudienmethode zur radiologischen Altersbestimmung von menschlichen radiologischen Leserinnen und Lesern eingeführt.

Der mittlere absolute Fehler (MAE) der Vorhersagen des Deep-Learning-Modells für das chronologische Alter betrug 1,65 Jahre, und der höchste beobachtete absolute Fehler lag bei 6,40 Jahren für Frauen und 7,32 Jahren für Männer. Die Ungenauigkeit in diesen Fällen mit hohem Fehler konnte jedoch auf Norm-Varianten oder pathologische Störungen zurückgeführt werden. Der mittlere absolute Fehler (MAE) der Schätzung der menschlichen Leserinnen und Leser betrug 1,84 Jahre, und der höchste berechnete absolute Fehler lag bei 3,40 Jahren für Frauen und 3,78 Jahren für Männer. Diese Ergebnisse zeigen, dass der entwickelte Deep-Learning-Ansatz für die kontinuierliche Altersvorhersage auf CT-Volumina der Klavikula die Genauigkeit der Altersschätzung der menschlichen Leserinnen und Leser im Durchschnitt übertrifft.

Zusammenfassend zeigt diese Arbeit verschiedene Machbarkeitsnachweise für maschinelles Lernen, die sich mit zwei klinischen Problemen in der Radiologie befassen: Darmkrebs-Screening mit CT-Kolonographie und radiologische Altersbestimmung auf der Grundlage der Verknöcherung des Schlüsselbeins in der CT. Die Ansätze lösten erfolgreich anspruchsvolle Probleme, die bei der konventionellen bildgebenden Diagnostik sonst nur schwer zu bewältigen sind.

1 | Introduction

Machine learning has been part of radiology for decades and already demonstrated successful applications in the 1990s [1–5]. Research in the radiology domain progressed alongside machine learning in general, which gained momentum with the introduction of the backpropagation algorithm for training neural networks in 1986 [6]. However, it was the groundbreaking performance of the convolutional neural network (CNN) AlexNet [7] in the 2012 ImageNet Challenge [8] that sparked the development of the countless machine learning applications existing today [9]. The field has since witnessed a trend of ever larger models that can solve increasingly complex tasks [10]. The rise of successful machine learning applications was made possible by the availability of previously missing key ingredients: large structured datasets with labeled examples to learn from and powerful computing resources for model training [5]. Especially deep learning benefitted from evolving multi-core GPUs designed for parallel matrix multiplications [7]. The exceptional opportunities demonstrated by radiomics in 2014 [11] additionally increased the interest of radiology departments worldwide in machine learning and its potential for medical image analysis.

A particular promise of machine learning to radiology is to push the boundaries of traditional imaging diagnostics and overcome challenges where traditional methods reach their limits. On that basis, this work investigated two clinical problems from two different areas of radiology and demonstrates machine learning approaches that surpass classical methods. First, the differentiation of premalignant from benign colorectal polyps in computed tomography (CT) colonography with machine learning [12, 13]. Second, improving the accuracy of radiological age assessment based on clavicle ossification in CT by enabling continuous age predictions using deep learning [14, 15].

CT colonography Colorectal cancer is one of the three leading causes of cancer-related death in industrialized countries [16]. It originates mostly from adenomatous polyps that slowly develop into cancer over the course of several years [17]. Early detection and removal of these premalignant adenomatous polyps can significantly reduce the incidence and mortality of colorectal cancer [18]. CT colonography is a non-invasive screening method for colorectal cancer and can detect polyps reliably

[19]. However, there is currently no definite way to differentiate between benign and precancerous polyp characters [12]. Instead, polyp size is used as a surrogate indicator and the resection of colorectal polyps larger than 6 mm is suggested [20, 21]. To address the need for polyp differentiation, this work introduces two machine learning approaches for the classification of colorectal polyps based on CT colonography scans [12, 13]. The goal is to enable individual risk stratification and therapy guidance after CT colonography examinations.

In the first approach, a random forest classification model is trained to predict the polyp character based on radiomic image features, calculated from CT colonography scans using manually annotated polyp segmentation masks [12]. The random forest is an ensemble of decision trees, trained on bootstrap resamples of the training data and randomly selected feature subsets [22]. The feature-extraction technique radiomics [23] is used to transform CT colonography scans into a vector of quantitative imaging biomarkers that can be processed by the random forest. Radiomics is often paired with random forests because they work well with high-dimensional inputs, provide error estimates and allow for feature importance analysis [22]. Thereby random forests give insight into the decision making process, which is particularly valuable in a sensitive environment like medicine. Additionally, annotated datasets for intricate medical findings, such as lesion character of small colorectal polyps, are often too small and inadequate for the development of more complex deep learning models.

In addition to the radiomics approach, this work also investigates a CNN for polyp classification based on CT colonography images [13]. CNNs are a particular group of deep learning algorithms that can directly map image inputs to clinical endpoints without relying on segmentation masks for classification [24]. These networks use convolutional units with a receptive field of view that are shifted across the input data to calculate features [25], and have been widely successful in image- and signal-processing [26]. Additionally, gradient-visualization techniques enable the highlighting of regions in the input image that are potentially important for model predictions, offering an intuitive form of model interpretability to radiologists [27]. The manual polyp segmentation in the radiomics approach is time-consuming and expensive and has high inter-reader variability [28], which makes it impracticable for the integration into everyday clinical workflows. Deep learning on the other hand provides the basis for a fully automated CT colonography evaluation, as CNN polyp classification could be combined with already established computer-aided detection algorithms for polyp detection [29, 30].

Despite these advantages, deep learning in radiology faces certain challenges. One major obstacle is the shortage of large, structured datasets with accurate ground-truth labels for models to learn from. In the medical domain, data annotation typically relies on human experts, which makes it a time-consuming and expensive process. Consequently, only a fraction of the vast amount of medical images stored in hospital databases worldwide is sufficiently labeled to address narrowly focused

medical research problems. Existing datasets are often too small to even fully exhaust the capabilities of established methods like CNNs. State-of-the-art deep learning techniques such as transformers [31] or diffusion models [32] require even more data, which dramatically limits their applicability in radiology.

Radiological age assessment A specific medical imaging problem where the availability of labeled data is less of an issue is radiological age assessment, a method at the intersection of radiology and forensic medicine to estimate the chronological age in living subjects based on radiographs or CT scans. The ground-truth label age is a parameter that is recorded for every patient prior to an examination in a hospital and is easily accessible in the picture archiving and communication system (PACS).

Age is an essential part of a person's identity, especially for a child, which by definitions of the United Nations and the European Union is any person below the age of 18 [33, 34]. Age determines the relationship between the state and the individual, and changes in age can trigger the acquisition or loss of rights and obligations [35]. When a person's age is unknown or in serious doubt, a state may need to assess the age, e.g., to determine whether they are an adult or a child.

One approach to age assessment is the radiological examination of the ossification status of the medial clavicular epiphyseal cartilages in CT scans [36]. In these methods, clinicians assess a clavicle ossification stage using defined criteria for differentiating between phases of skeletal maturation. The age of the individual in question is then assumed to be similar to a case group with similar skeletal maturation features. One inherent problem with that approach is the limited accuracy due to the limited number of ossification stages that can be assessed with the human eye, leading to a small available set of discrete age estimates. To address this issue, this work investigates mapping thoracic CT scans to chronological age using a CNN in order to enable continuous age prediction [15]. Additionally, we compare the CNN approach to a favorable estimate of the human-reader performance for the widely acknowledged reference-study method of Kellinghaus et al. [36–39]. The goal is to improve age assessment accuracy compared to the reference-study method.

The training process for diagnostic deep learning models benefits from inputs that are cropped to the region of interest (ROI) containing information relevant to solving the problem. Therefore, it is advisable to first localize the medial clavicular epiphyseal cartilages within the thoracic CT scans before training the age assessment CNN [40]. However, localization by human experts would be time-consuming and expensive, which would obstruct the creation of large datasets [41]. This work demonstrates an object detection approach for the automated localization of the clavicles in the thoracic CT scans used for age assessment [14]. The goal is to remove the human annotation bottleneck and enable large labeled datasets that facilitate the training of suitable deep learning models to potentially improve radiological age assessment.

2 | Background

2.1 X-ray Computed Tomography

X-ray imaging is not only the most common form of medical imaging but also the oldest. After Wilhelm C. Röntgen had discovered X-rays as a "new type of radiation" in 1895 [42], they were quickly utilized to acquire the first documented radiographic image, a projection of the hand of his wife Anna B. Röntgen (Figure 2.1) [43]. The energy of X-ray photons is high enough that they can penetrate human tissue and enables imaging inner structures of a human.

Figure 2.1: First documented X-ray showing the hand of Anna B. Röntgen. Rights: Deutsches Röntgen-Museum [43].

A particular medical imaging technique based on X-rays is called X-ray computed tomography (CT) (from Greek $\tau o \mu \eta$, *tome* 'section', and $\gamma \rho \alpha \phi \eta \iota \nu$, *graphein* 'to write') and can create three-dimensional reconstructions of an object and its interior. Allan M. Cormack first proposed the technique in 1963 as the *Representation of a Function by Its Line Integrals, with Some Radiological Applications* [44]. Following up on his idea, the object is irradiated with a beam of X-rays in order to acquire line integrals of the imaged quantity, the X-ray attenuation coefficient. In simple

radiographs or 2D X-ray projection images, the X-ray attenuating effects (e.g. tissue density) are superimposed and the final image contains no depth information. CT reconstructs this depth information by acquiring multiple projections from different angles, which enables the examination of patients in three dimensions. This extra information can reveal potentially important details in organs, joints, blood vessels, and more. Because of CT's impact on the field of medicine, Allan M. Cormack and Godfrey N. Hounsfield were awarded the 1979 Nobel Price in Physiology or Medicine[1] for *The development of computer assisted tomography.*

This section is a short introduction to the basic principles of X-ray CT. It gives an overview of X-ray generation (2.1.1), X-ray interaction with matter (2.1.2), X-ray detection (2.1.3) and finally image acquisition (2.1.4), including image reconstruction. If not stated otherwise, the information in this section is based on the textbooks *Computed Tomography* by Buzug [45] and *Strahlenschutz für Röntgendiagnostik und Computertomografie* by Grunert [46].

2.1.1 X-ray Generation

X-rays are high-energy electromagnetic waves and can be generated by decelerating fast electrons in an anode material with positively charged atoms, e.g. tungsten, molybdenum, or copper. The wavelength of X-rays is roughly between 10^{-8} m and 10^{-13} m.

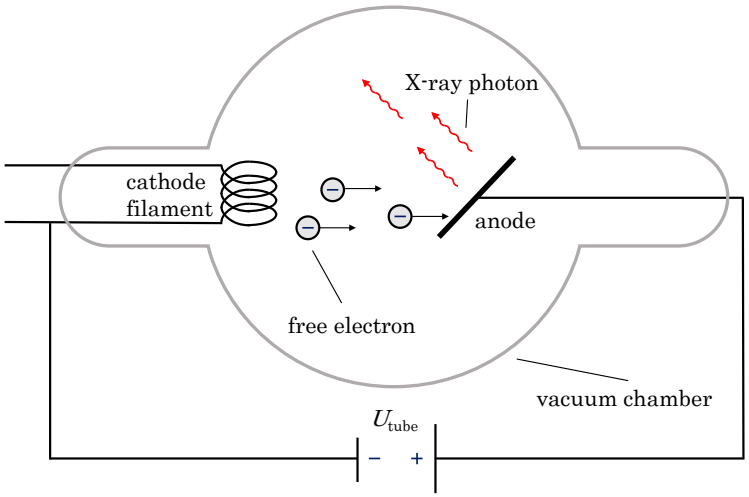

Figure 2.2: Schematic illustration of a simple X-ray tube. The cathode filament is heated until the electrons bound inside become free and are accelerated in the electromagnetic field controlled by the tube voltage U_{tube}. When the accelerated electrons rapidly decelerate in the anode, X-rays are generated as Bremsstrahlung.

The first step to generate X-rays is creating fast free electrons. This is typically

[1] https://www.nobelprize.org/prizes/medicine/1979/summary/

performed with an X-ray tube (Figure 2.2), which consists of a cathode and an anode inside a vacuum chamber. On the cathode end is a filament, e.g. made of thoriated tungsten (melting point = 3400 °C), that is heated to the point where the kinetic energy of the electrons in the filament is high enough to overcome the binding energy. This allows thermal electrons to escape the cathode filament and become free electrons which can be accelerated by an electromagnetic field. That field is created by applying a so-called tube voltage U_{tube} between the anode and cathode, which also determines the kinetic energy of the electron neglecting relativistic effects

$$eU_{\text{tube}} = \frac{1}{2} m_e v^2 \quad . \tag{2.1}$$

Typical tube voltages found in medical X-ray imaging are between 25 kV and 125 kV.

To generate X-rays, the fast accelerated electrons are slowed down abruptly in the anode material. This deceleration is a combination of different processes and leads to the emission of a continuous spectrum of Bremsstrahlung photons that is superimposed by photons from characteristic emission, the Auger process, and direct electron-nucleus collisions (Figure 2.3).

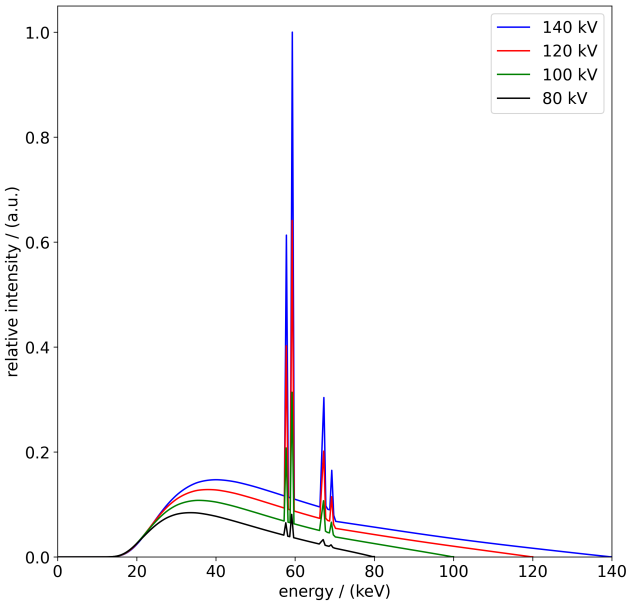

Figure 2.3: X-ray spectra of a tungsten anode for electrons accelerated in an electric field with voltages from 80 to 140 keV. The anode was at a 10° angle and the spectrum was filtered by 2 mm of aluminum. The spectra were modeled with the open-source toolkit SpekPy (version 2.0.8) [47].

Electron interaction with matter During deceleration, almost all of the traveling electrons' kinetic energy is turned into heat (roughly 99 %) and taken up by the

anode. However, while the electrons are stopped by the Coulomb fields of the atoms, a small part of their energy is lost to Bremsstrahlung and emitted as a continuous X-ray spectrum (Figure 2.3). The closer the electrons get to the nuclei of the atoms in the anode material, the more energy they lose during deceleration and the higher the frequency of the Bremsstrahlung photons get. Typically, a single decelerated electron causes the emission of multiple photons. Very rarely, an electron can also convert its entire energy into a single photon in a process called direct electron-nucleus collision. These photons represent the upper end of the X-ray spectrum and have the maximum energy

$$E_{\max} = h\nu_{\max} = eU_{\text{tube}} \quad . \tag{2.2}$$

Therefore, the tube voltage determines the upper limit of the energy interval of the generated X-ray spectrum, and the intensity of the X-ray spectrum is controlled by the total number of electrons, i.e. the anode current.

The continuous Bremsstrahlung spectrum is superimposed by so-called characteristic emission (Figure 2.3). When an electron ionizes an atom of the anode material by removing an inner electron, an electron of one of the higher shells fills the vacant position. Consequently, photons with quantized energy are emitted that can be seen as distinct intensity lines of high photon intensity in the X-ray spectrum and are characteristic of each anode material. Another interaction is the Auger process. It is seen as a non-radiation process and describes a scenario where a photon, that otherwise would have been characteristic emission, is absorbed by the atom and another electron is emitted instead.

2.1.2 X-ray Interaction with Matter

X-rays have a good capability to penetrate matter, including human tissue. During penetration, the intensity of an X-ray beam decreases exponentially due to different absorption and scattering effects. In CT the goal is to measure this attenuation of X-ray beam intensity caused by the patient from different angles and use the information to reconstruct the interior structure of the patient. The most important photon-matter interaction mechanisms are briefly described in this section: Rayleigh scattering, photoelectric absorption, Compton scattering, and pair production.

Rayleigh scattering Rayleigh (or Thomson) scattering is an elastic scattering process. No energy is transferred and only the photon direction is changed. It requires the diameter of the scattering nucleus to be small compared to the photon wavelength. In the classical model of Rayleigh scattering, an incoming photon causes bound electrons of an atom in the penetrated material to oscillate. This oscillation creates a dipole, which then radiates a photon in an arbitrary direction. The Rayleigh scattering cross-section is given by

$$\sigma_{\text{Thomson}} \propto \frac{\omega^4}{(\omega^2 - \omega_0^2)^2} \quad , \tag{2.3}$$

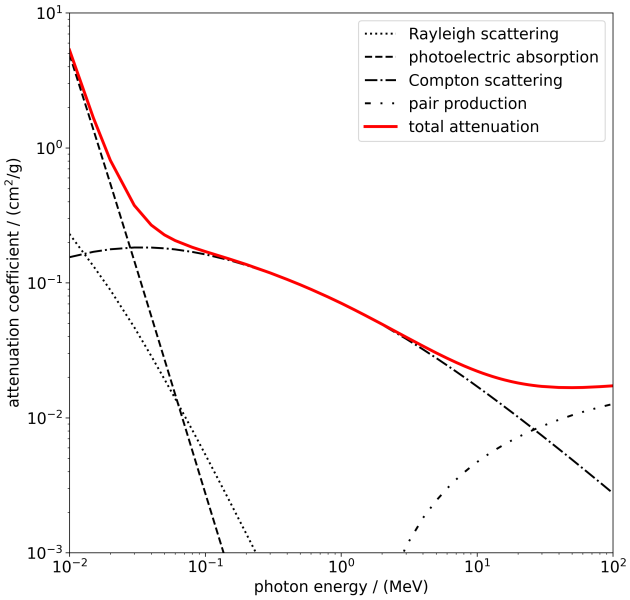

Figure 2.4: Attenuation coefficient for photons in water (H_20) for incident photon energies typically found in medical imaging. Major effects contributing to the total attenuation are Rayleigh scattering, photoelectric absorption, Compton scattering, and pair production. Photon cross sections were calculated with the XCOM: Photon Cross Sections Database [48].

where ω_0 is the natural frequency of the bound electrons [45]. This scattering process plays an important role in photon attenuation at low energies where $\omega < \omega_0$. However, at high energies other competing processes become dominant.

Photoelectric absorption Photoelectric absorption is another photon-matter interaction, in which a photon is entirely absorbed by an atom. It can occur when the binding energy of the electron is less than the photon energy. As a result of the photon absorption, the atom is ionized. An electron is kicked off the atom and uses the difference between photon energy and binding energy as kinetic energy to travel as a photoelectron. This process is known as the photoelectric effect [49]. The absorption coefficient has been demonstrated to depend on the atomic number Z of the penetrated material and the frequency of the incident photon ω:

$$\sigma_{\text{Photo}} \propto \frac{Z^4}{\omega^3} \quad . \tag{2.4}$$

The vacancy left behind by the photoelectron is filled with an electron from a higher shell or the electron band. During this recombination process, a photon is emitted and causes characteristic X-ray fluorescence. In case the energy of the emitted photon is high enough to remove other electrons, the previously described Auger process (Section 2.1.1) is triggered.

Compton scattering Compton scattering is an inelastic scattering process in which an incoming photon collides with a quasi-free electron. The amount of energy that is transferred from the photon to the electron depends on the scattering angle θ of the photon. The wavelength shift is given by

$$\Delta\lambda = \frac{h}{m_e c}(1 - \cos\theta) \quad . \tag{2.5}$$

Significant parts of the scattered photons are scattered at an angle $\theta > 90\,^\circ$ and are subsequently traveling backward. Because it is an interaction with quasi-free electrons, e.g. weakly bound valence electrons in the outer shell, the cross-section does not depend on the atomic number of the penetrated material.

Pair production The final interaction mechanism to be discussed in this section is pair production. Photons with an energy above $2 \times 511\,\text{keV}$ have a chance to create an electron-positron pair inside the Coulomb field of a nucleus or an electron. When the positron collides with an electron, they annihilate and two photons are emitted in opposite directions.

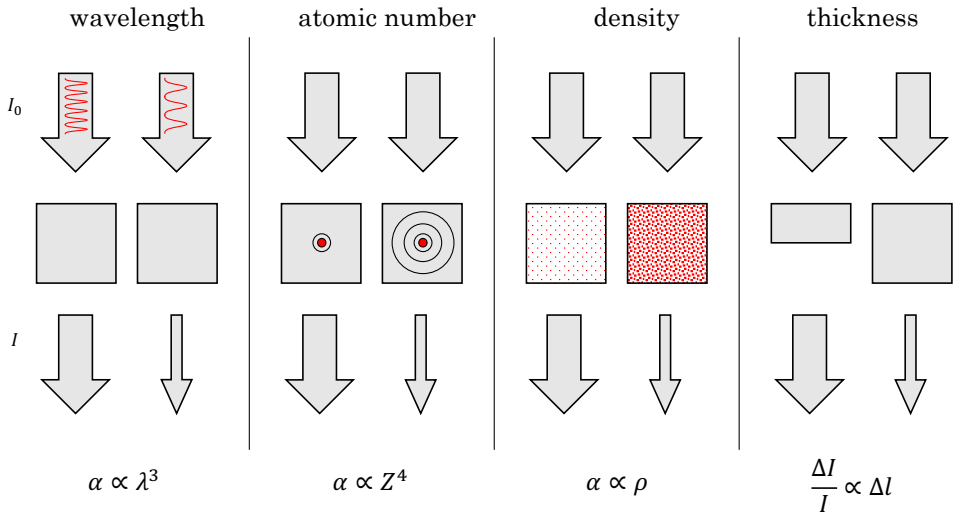

Figure 2.5: Schematic illustration of the simplified relationship between X-ray attenuation and incident photon properties, as well as material properties. The attenuation is higher for longer wavelengths, higher atomic numbers, higher material density, and thicker materials. Adapted from [50] and [45].

In summary, the attenuation of X-ray photons in matter depends on the wavelength of the incident photons, as well as the atomic number, mass density, and thickness of the penetrated material. This behavior is schematically illustrated in Figure 2.5.

Beer-Lambert Law A common equation used to calculate photon beam attenuation is the Beer-Lambert law (Equation 2.6). It describes the intensity I of a photon

beam after it has traveled a distance x through an object with the material-dependent attenuation coefficient μ as:

$$I(x) = I_0 e^{-\mu x} \quad . \tag{2.6}$$

The Beer-Lambert law makes several simplification assumptions. First, it is based on a classical scattering model and no quantum effects are considered. Second, the penetrated object is assumed to be homogenous and all beam intensity-decreasing interactions are summarized in a single, constant attenuation coefficient μ. Finally, the law only holds for a pencil-beam geometry of monochromatic photons, where each scattered photon is fully removed from the beam.

In reality, the patients radiated in medical examinations are not homogenous, but are complex compositions of tissues, organs, and blood. Therefore, the attenuation is spatially dependent, $\mu(x)$. Additionally, the photon-matter interactions described earlier depend on the incident photon energy. Because the energy of the photons decreases while traveling through the patient, the attenuation coefficient is also energy dependent, $\mu(E)$. Considering both effects, the Beer-Lambert may be extended to:

$$I(x) = \int_0^E I_0(E) e^{-\int_0^l \mu(E,x)\mathrm{d}x} \mathrm{d}E \quad . \tag{2.7}$$

However, the energy dependence is typically neglected in CT image reconstruction, which leads to so-called beam hardening artifacts in the reconstructed image.

2.1.3 X-ray Detection

X-ray detection in CT scanners is an indirect process and relies on photon-matter interaction products. The efficiency of detection is determined by two factors: geometric efficiency and quantum efficiency. Geometric efficiency is the relative active detector area. It is calculated by dividing the active area by the total exposed area, which includes parts like the antiscatter grid (Figure 2.6) where no photons can be detected, and that thus lower geometric efficiency. Quantum efficiency refers to the probability that a photon hitting the detector will be detected and depends on the incident photon energy, as well as the detector material. There are different technologies for detecting X-rays in CT scanners, but this brief overview focuses on energy-integrating detectors (EIDs). For EIDs, there are two basic detector types: solid-state scintillators and high-pressure gas ionization detectors.

Today, most detectors in CT scanners are scintillator detectors (Figure 2.6), which consist of a scintillator material and a photon detector. When a high-energy X-ray photon hits the scintillator material, it is converted into long-wavelength scintillation light, which is then converted into an electronic signal by a photodiode. This signal is stored in a capacitor, allowing for the integration of energy from multiple photons. After a certain time, the energy stored in the capacitor is read out and used as the intensity signal for the respective detector pixel.

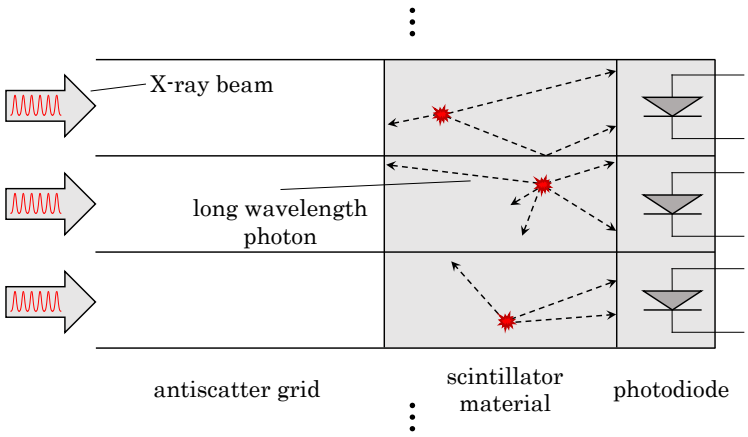

Figure 2.6: Schematic illustration of a solid-state scintillator X-ray detector. Adapted from [45].

Another recently developed type of detector is the photon-counting detector (PCD), which offers several advantages over EIDs such as lower noise and reduced required dose. However, the clinical availability of PCDs is still limited [51].

2.1.4 Image Acquisition

This section briefly introduces CT image acquisition and reconstruction. The image acquisition process is described using a parallel pencil beam geometry and 2D data for simplicity. However, CT scanners in clinical practice today operate slightly differently and typically use a spiral and 3D cone-beam geometry.

The goal of CT is to determine the object function $f(x, y)$, which in this case is the spatially dependent attenuation coefficient $\mu(x, y)$. It describes how much X-ray radiation is absorbed by the patient's body at each point. The coordinate system (x, y) is the fixed patient coordinate system. The X-ray source emits a single, needle-like X-ray beam and is moved linearly on the source (or detector) axis ξ in small steps. The detector is moved along with the source respectively. The X-ray attenuation caused by the patient at each point ξ is measured, yielding a one-dimensional projection or radiograph for each projection angle γ. The acquisition of a single projection is illustrated schematically in Figure 2.7.

Mathematically, the attenuation of a single pencil beam is given by the projection integral p and depends on the material depth $\Delta\eta$, the position of the X-ray source (or the detector) ξ, and the projection angle γ:

$$p_\gamma(\xi) = \int_0^s \mu(\xi, \eta)\mathrm{d}\eta \tag{2.8}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\delta(x\cos(\gamma) + y\sin(\gamma) - \xi)(\xi, \eta)\mathrm{d}x\mathrm{d}y \quad . \tag{2.9}$$

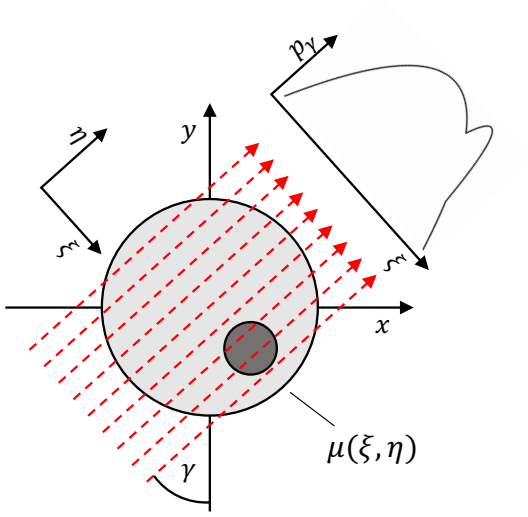

Figure 2.7: Schematic illustration of the acquisition of a single projection in CT. A number of parallel pencil beams penetrate the object with the spatially dependent attenuation coefficient $\mu(\xi, \eta)$ at an angle γ. A detector measures the attenuated beam intensity behind the object at each position ξ, yielding the projection (or one-dimensional radiograph) $p_\gamma(\xi)$.

The coordinate system (ξ, η) describes the rotating scanning system. Projection data is acquired by measuring projection integrals from parallel X-ray beams for angles γ between 0 and 180 °. Angles above 180 ° project the X-ray path back through the object and provide no additional information. A full set of projections $p_\gamma(\xi)$ is called the Radon transform or sinogram of the image.

Fourier slice theorem So far, the process of measuring projections $p_\gamma(\xi)$ has been described. However, the goal of CT is to determine $f(x, y)$. The so-called Fourier slice theorem enables a direct method to reconstruct an image of $f(x, y)$ based on $p_\gamma(\xi)$. The theorem states that the one-dimensional Fourier transform of the projection $p_\gamma(\xi) \longrightarrow P_\gamma(q)$ can be identified with a radial line in the Cartesian Fourier space $F(u, v)$ of the object at the projection angle γ of the corresponding measurement:

$$F(u, v) = P_\gamma(q) \quad . \tag{2.10}$$

This enables recovering the object $f(x, y)$ in spatial coordinates by applying an inverse two-dimensional Fourier transform to $F(u, v)$, which in turn can be derived from the Radon transform of the object $f(x, y)$, i.e. the measured projections. The relationship between the object space, projection space (Radon space), and Fourier space is summarized in Figure 2.8.

The steps to reconstruct a two-dimensional CT image slice using the Fourier slice theorem are as follows:

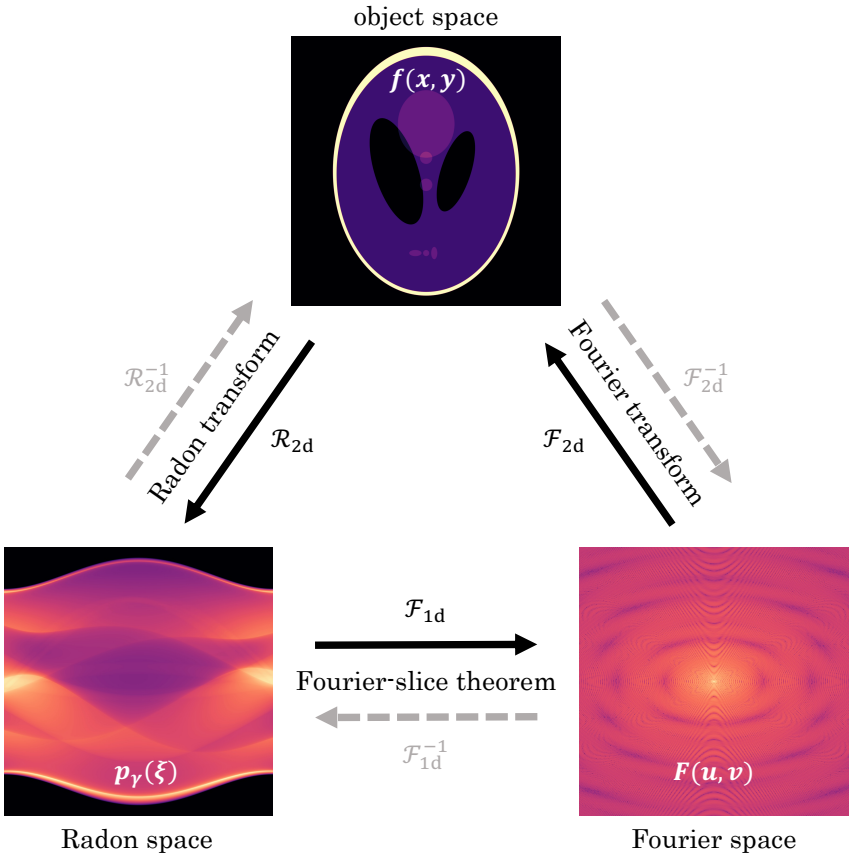

Figure 2.8: Illustration of the relationship between object space, Radon space, and Fourier space. According to the Fourier slice theorem, the Fourier transform of the object's Radon transform is equal to the inverse Fourier transform of the object. The object $f(x, y)$ is the Shepp-Logan phantom [52].

1. Acquire the Radon transform $p_\gamma(\xi)$ of the object $f(x, y)$ through measuring X-ray projections
2. Calculate the Fourier transform of the Radon transform, $p_\gamma(\xi) \longrightarrow P_\gamma(q)$
3. Apply the Fourier slice theorem to transform P into the Fourier transform of f, $P_\gamma(q) \longrightarrow F(u, v)$
   - This step requires a change in coordinates from polar coordinates (q, γ) to Cartesian coordinates (u, v) through the substitution

$$u = q\cos(\gamma) \tag{2.11}$$
$$v = q\sin(\gamma) \quad . \tag{2.12}$$

– In practice, this requires the spectral space (u, v) to be filled densely with data points by measuring projections at a high number of angles γ with parallel pencil beams at many positions ξ

4. Calculating the inverse Fourier transform of F, $F(u, v) \longrightarrow f(x, y)$

The direct image reconstruction workflow is illustrated schematically in Figure 2.9.

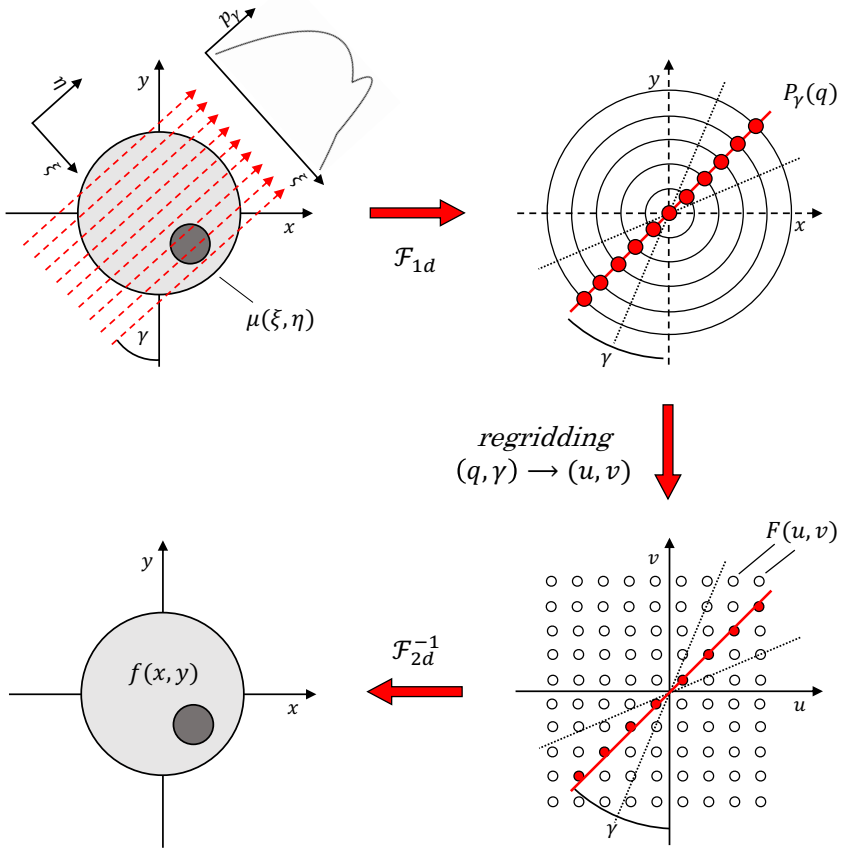

Figure 2.9: Schematic illustration of a direct image reconstruction process in CT. First, the Radon transform $p_\gamma(\xi)$ of the object $f(x, y)$ is acquired through measurement. Second, the Radon transform is transformed into Fourier space to obtain $P_\gamma(q)$. Next, the measured data points in the radial (γ, ξ) space are regridded onto the Cartesian grid (u, v), which transforms $P_\gamma(q)$ into $F(u, v)$. Finally, the inverse Fourier transform of $F(u, v)$ yields the object $f(x, y)$.

Unfortunately, this method is not able to accurately recover the object function $f(x, y)$ in clinical practice. Because of dose considerations and technical limitations, a CT scanner is limited in the number of projections that can be measured. The finite number of data points in (q, γ) from the measurements have to be regridded to fill the Cartesian (u, v) space, which leads to interpolation errors. As illustrated in

Figure 2.10, the density of measured spectral data in the (u, v) space decreases for higher frequencies.

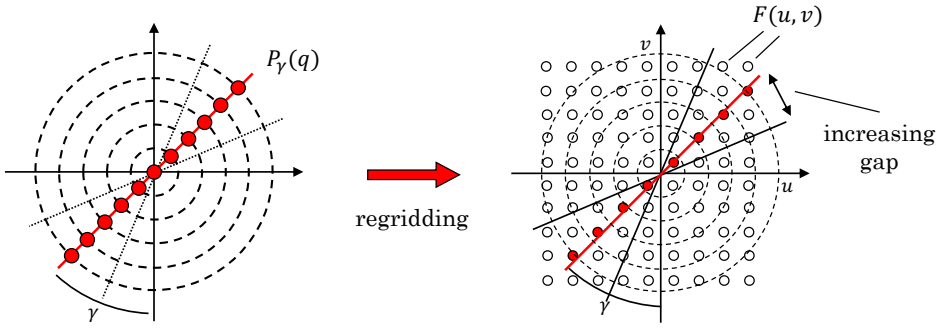

Figure 2.10: Illustration of regridding measured data points from the polar Radon space (q, γ) to the Cartesian Fourier space (u, v). The measured data points on the large concentric circles do not always overlap with the points in the (u, v) grid. In order to fill the (u, v) space, the measured data has to be interpolated. The further away from the center, the larger the distance between the measured projections in the (u, v) space and the higher the interpolation error.

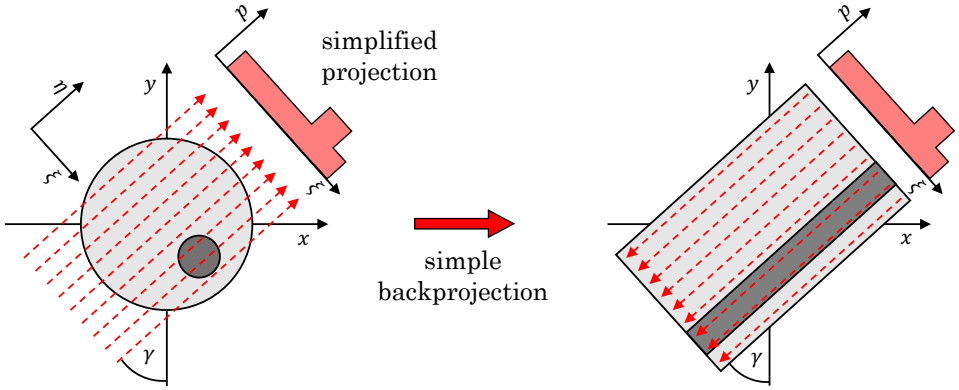

Figure 2.11: Simple backprojection of the (simplified) projection $p(\xi)$ measured at angle γ.

Filtered backprojection A particular reconstruction technique found in clinical practice is called *filtered backprojection*. The idea is that the image can be reconstructed by projecting the measured projection profiles $p_\gamma(\xi)$ back for each angle γ. This form of simple (non-filtered) backprojection is illustrated in Figure 2.11 and can be modeled by

$$g(x, y) = \int_0^\pi p_\gamma(\xi) \mathrm{d}y \quad . \tag{2.13}$$

However, the simple backprojection in Equation 2.13 does not result in the object function $f(x, y)$ and the spatial distribution of the attenuation coefficient $\mu(x, y)$, respectively. The problem is that the projection profile $p_\gamma(\xi)$ is a non-negative function and the simple backprojection smears back non-negative values over the entire image, even outside the actual object.

One solution is filtered backprojection, where the projection signal $P_\gamma(q)$ is high-pass filtered in Fourier space by multiplying $P_\gamma(q)$ with $|q|$ to achieve linear weighting of each frequency. Mathematically, the high-pass filter can be derived as a result of the coordinate transformation from Cartesian to polar coordinates using the Fourier slice theorem. Starting from the image $f(x, y)$ expressed as the inverse Fourier transform of $F(u, v)$

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{2\pi i(xu+\gamma v)}\mathrm{d}u\mathrm{d}v \quad , \tag{2.14}$$

the Fourier transform $F(u, v)$ has to be expressed in polar coordinates using the substitution from Equation 2.11. Together with the Fourier slice theorem in Equation 2.10, this ultimately results in the high-pass filtered backprojection:

$$h_\gamma(\xi) = \int_{-\infty}^{\infty} P_\gamma(q)|q|e^{2\pi iq\xi}\mathrm{d}q \quad . \tag{2.15}$$

Details of the mathematical derivation can be found in [45]. Finally, the attenuation function can be calculated from a set of filtered backprojections as follows:

$$\mu(x,y) = f(x,y) = \int_0^{\pi} h_\gamma(\xi)\mathrm{d}\gamma \quad . \tag{2.16}$$

Besides filtered backprojection, different reconstruction techniques exist, including algebraic and iterative approaches.

Hounsfield units In CT, the measured X-ray attenuation values μ are usually represented as gray values and are a representation of the physical material properties. High values of the attenuation coefficient μ relate to a high density or high atomic number of the medium. This is a relevant difference compared to magnetic resonance imaging (MRI), where the relation between gray values and physical properties is more complex, as it depends on many different parameters and scanning protocols. In clinical practice, X-ray attenuation values are expressed as dimensionless CT values in Hounsfield units (HU). To this end, they are transformed as follows:

$$\text{CT-Value} = \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}}} 1000\,\text{HU} \quad . \tag{2.17}$$

On the Hounsfield scale, the value -1000 HU is assigned to air (-1024 HU $\widehat{=}$ vacuum) and 0 HU to water. Therefore, CT values can be used to identify organs and tissue types (see Table 2.1) and diagnose pathologies, e.g., when the intensity distributions of CT values show abnormalities or saliences.

tissue	attenuation value / (HU)
air	-1000
bone	>250
fat	-200 to 50
lung	-900 to -500
parenchyma	0 to 100
tumor	20 to 50
water	0

Table 2.1: Approximate attenuation value windows in Hounsfield units of selected tissues. Data taken from [45, 53, 54].

2.2 CT Colonography

CT colonography is a non-invasive X-ray-based screening method for colorectal cancer. This section gives a brief overview of colorectal cancer, CT colonography screenings, and the strengths and weaknesses of CT colonography.

2.2.1 Colorectal Cancer

Colorectal cancer is among the three leading causes of cancer-related death in industrialized countries for both men and women [16]. Most colorectal cancers originate from adenomatous polyps, which develop slowly over several years into colorectal cancer [17]. Early detection and removal of these premalignant adenomatous polyps can significantly reduce the incidence and mortality of colorectal cancer [18]. In the early stages of colorectal cancer, symptoms are often nonspecific or nonexistent, making screening methods such as optical colonoscopy essential for cancer prevention [55]. However, the participation rates in colonoscopy screenings are only around 15 to 20 % [56].

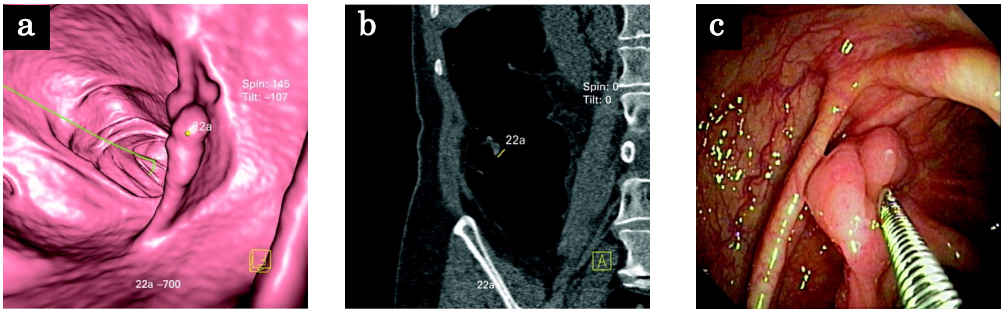

Figure 2.12: (A) CT colonography image of a 2.2 cm sessile polyp in the ascending colon in a 72-year-old asymptomatic female. (B) CT image of the same polyp. (C) Optical colonoscopy image of the same polyp. Adapted from [19]. Reuse permitted by BMJ Publishing Group Ltd. (License Number 5552401418559).

2.2.2 CT Colonography Screening Method

A particular screening method that has emerged over the last two decades is CT colonography, also known as virtual colonoscopy. A CT colonography screening involves a low-dose abdomen CT to acquire detailed three-dimensional reconstructions of the colon. Additionally, the colon is cleansed and inflated with CO_2 prior to image acquisition. This removes fecal matter that may obstruct the view, which allows for clear images of the colon and helps to reveal detailed structures of the intestinal wall. In addition to regular multi-planar CT image reconstruction, CT colonography includes a virtual three-dimensional reconstruction of the colon, allowing for polyp identification similar to optical colonoscopy. Figure 2.12 shows a virtual colon reconstruction, a common CT reconstruction, and an optical colonoscopy image of the

same polyp. The sensitivity of CT colonography is comparable to optical colonoscopy for detecting colorectal polyps that are 6 mm or larger in size [19]. CT colonography has an advantage in cases of complex colon anatomies as it enables visualization of parts of the colon that may not be examined with optical colonoscopy [57]. Screening programs using CT colonography show about 10 % higher participation rates compared to optical colonoscopy programs [58, 59].

However, while CT colonography can detect polyps, it cannot distinctively differentiate between benign and precancerous polyps. Currently, size serves as a proxy indicator for malignancy. Guidelines from the *United States Multi-Society Task Force on Colorectal Cancer*, the *European Society of Gastrointestinal Endoscopy*, and the *European Society of Gastrointestinal and Abdominal Radiology* recommend the resection of colorectal polyps that are 6 mm or larger in size [20, 21]. Reliable differentiation between benign and precancerous polyps is crucial for individual risk stratification and guidance in determining the appropriate therapy.

2.3 Radiological Age Assessment

Radiological age assessment is a method in the field of forensic medicine to estimate the chronological age of a person by analyzing physical development and skeletal maturation using medical images. This section is an introduction to the role of age in modern society and age assessment methodology, including its weaknesses. The information in this section is largely based on the *EASO Practical guide on age assessment* by the European Union Agency for Asylum (EUAA) [35] and the publication *Forensic Age Estimation: Methods, certainty, and the law* by Schmeling et al. [36].

2.3.1 Age and Society

Age is an essential part of a person's identity, as it rules the relationship between the individual and the state. Changes in age can trigger the acquisition or loss of rights and obligations concerning emancipation, employment, criminal responsibility, sexual relation, and consent for marriage or military service [35]. This is particularly relevant for children. The United Nations Convention on the Rights of the Child (CRC) (Article 1) and the European Union[2] define a child as any person below the age of 18 [33, 34].

Additionally, the CRC lists certain age-related key obligations for states and authorities that include registering the child after birth, respecting the right of the child to preserve his or her identity, and speedily re-establishing his or her identity in case some or all elements of the child's identity have been deprived [35]. Consequently, a state may need to assess the age in cases where a person's age is unknown or there are substantiated doubts concerning the available age information, e.g., to determine whether they are an adult or a child. Authorities and courts can call every physician with sufficient expertise as an expert in order to conduct age assessments [36]. However, age assessments are typically performed by forensic physicians, radiologists, dentists, primary care physicians, and pediatricians [36].

The EUAA recommends that the least intrusive and most accurate method should be selected for the age assessment, gradually implementing (Figure 2.13) more invasive methods if deemed necessary [35]. Additionally, the EUAA suggests documenting the margin of error of the method applied. Methods involving potentially harmful radiation should only be applied as a last resort. In the best interest of the child, the EUAA outlines the following prioritization of age assessment methods [35]. First, non-medical methods should be applied, including further assessment of evidence regarding the age, an age assessment interview, and a psychological assessment. Second, radiation-free medical methods like dental observation, MRI scans, or assessment of physical development may be used. Last, and provided that there is a legal basis, medical methods involving potentially harmful radiation may be applied.

[2] EU acquis, Directive 2013/33/EU, Article 2(d)

These radiation-based methods include carpal (hand) X-ray imaging, collar bone X-ray imaging, and dental X-ray imaging. In Germany, the health of the individual is not necessarily the sole requirement for a legal justification for such examinations, as the anticipated benefit to the public from the applicable laws is also taken into consideration [36].

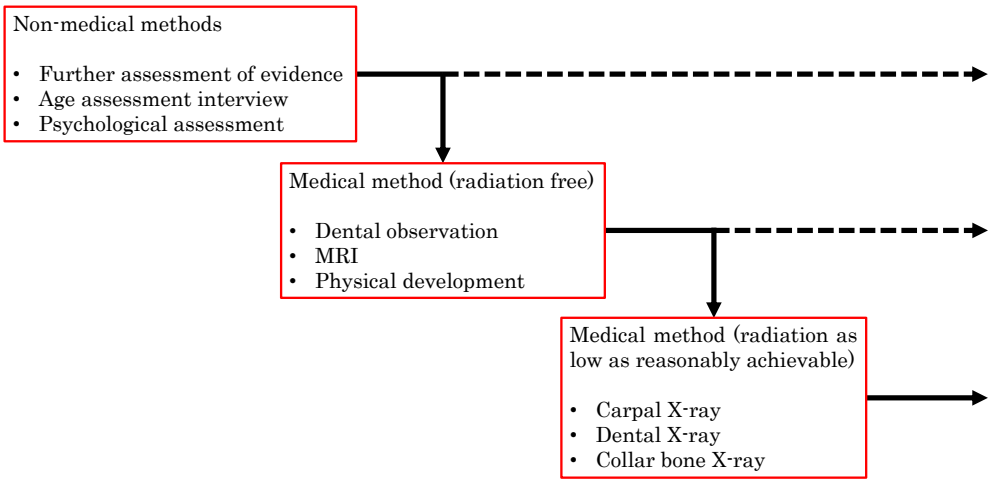

Figure 2.13: Gradual implementation of age assessment methods recommended by the EUAA. Information from [35].

2.3.2 Radiological Age Assessment Methodology

Typically, age assessment starts with an initial medical assessment in order to identify or rule out growth and developmental disorders. The Study Group on Forensic Age Diagnostics (AGFAD) (German *Arbeitsgemeinschaft für Forensische Altersdiagnostik*) of the German Society of Legal Medicine recommends beginning with an interview to take the medical history, followed by a physical examination to record the height, weight, and other characteristics of the individual [60]. Afterward, the radiological age assessment may be conducted. Age assessment is based on the known temporal progression of certain human development characteristics shared by everyone, including physical development, skeletal maturation, and dental development [36]. Starting from this, reference studies or atlases correlate the development of these characteristics with the known chronological age and sex of individuals or case groups. The most famous example is the Greulich and Pyle atlas [61] used for the determination of bone maturity from hand radiographs, e.g. to diagnose pediatric disorders.

Examinations for age assessment include hand radiographs, dental orthopantomograms (radiographs of the mandible, maxilla, and teeth), and thoracic CTs (Figure 2.14) [36]. Hands are evaluated with respect to the size, form, and ossification status of the epiphyseal plates [61]. In dental orthopantomograms, the evaluation focuses

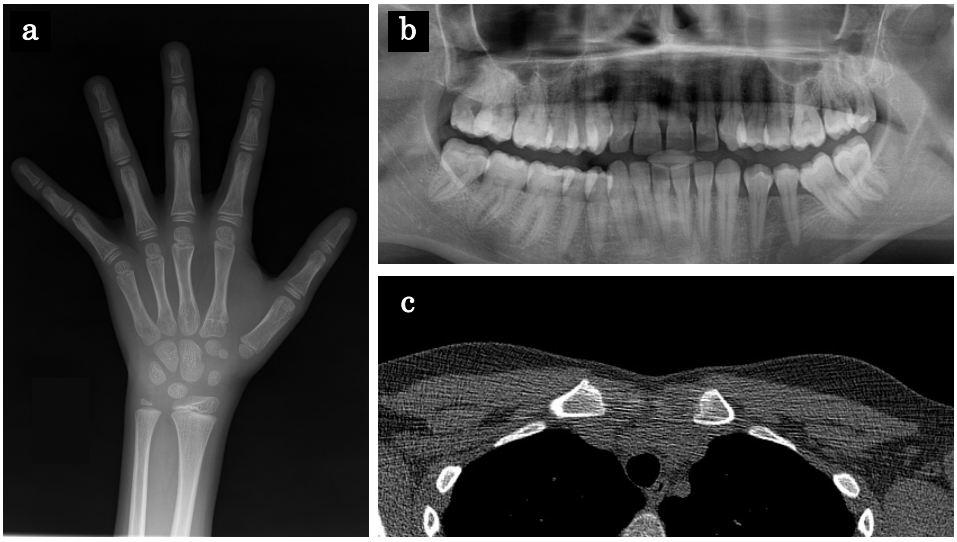

Figure 2.14: Typical X-ray-based examinations for age assessment. (a) Hand X-ray (8.0 year old male). (b) Dental orthopantomogram (unknown age and sex). (c) Axial slice from a thoracic CT scan (18.7 year old female).

on the eruption and mineralization of the third molars [62]. Thoracic CTs are used to assess the ossification status of the medial clavicular epiphyseal cartilages [37, 38]. The clavicular epiphysis is of particular interest. As the last maturing bone structure in the body, it allows age assessment not only for minors but also for young adults [63]. During all of these evaluations, the examined body parts are compared against atlases and reference studies. For instance, the age of the examined person may be assumed to be similar to a case group from a reference study that shows similar skeletal maturation.

In practice, the joint information from examinations of different body parts is often used to set upper and lower limits for the estimated age [36]. For instance, a state might want to assess whether a person is potentially a child, i.e. younger than 18 years, and requests a radiological age assessment. Next, a team of forensic pathologists, radiologists, and dentists examines the hands, molar teeth, and clavicles of the person in question. Typically, the estimated minimum age is reported following each examination, as well as the most likely age and the estimated maximum age, if available [36]. The results might look like the fictitious assessment shown in Figure 2.15. Here, the estimated age of the person in question was older than 15.9 years based on a hand X-ray, older than 17.5 years based on a dental X-ray of the molars, and between 16.6 and 22.4 years based on a clavicle CT, with the most likely age being 19.7 years. Concluding the three reports, it is possible that the examined individual is a minor, as no individual assessment indicates a minimum age above 18. Overall, the expected age of the person is estimated between 17.5 and 22.4 years.

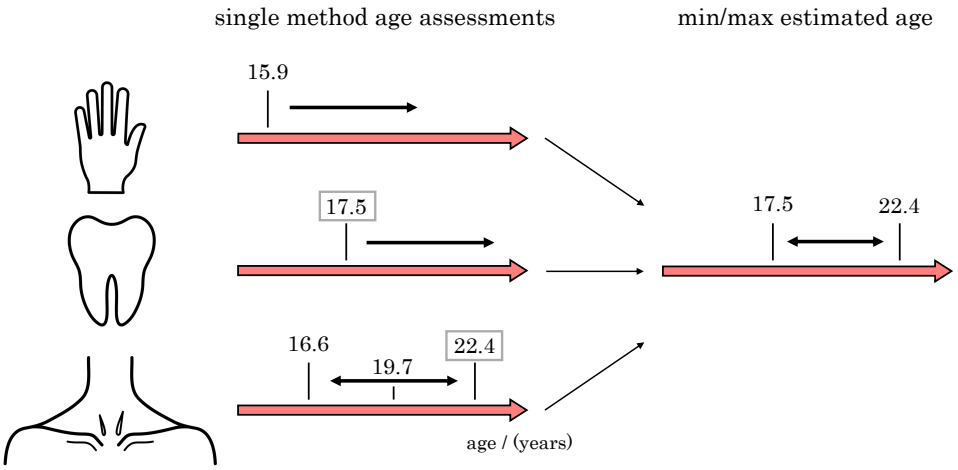

Figure 2.15: Schematic illustration of combining age assessment results from different examinations (hand, teeth, and clavicles) to estimate a minimum and maximum age. Icons from *thenounproject.com*, created by Icongeek26 (Hand #5761051), Alina Oleynik (Tooth #1092367), and Olena Panasovska (Collarbone #3292884).

2.3.3 Limitations of Radiological Age Assessment

Age assessment methods based on reference studies and atlases have inherent limitations. The complex relationship between skeletal development and chronological age poses an insurmountable natural accuracy barrier (bone age $\neq$ chronological age) [64]. Skeletal maturation depends on a variety of factors ranging from genetic predisposition to socio-economic status [65], which are difficult to account for during the evaluation of radiographs or CT scans described in Section 2.3.2. Moreover, case groups are typically examined at a single institution [36–39, 66], which limits diversity and introduces biases in the statistical analysis. Therefore, age assessment suffers from low accuracy, as well as intra- and inter-reader variability [67, 68]. In particular, the accuracy is limited due to the finite number of subgroups with known ages typically found in case groups. For instance, the well-established method of Kellinghaus et al. [37, 38] analyzes the ossification status of the sterno-clavicular joint. However, the examined case group consists of only 9 subgroups - 5 major ossification stages and 3 substages for stages 2 and 3, respectively - that a person in question can be compared against. Because of all these limiting factors, it is important to highlight that radiological age assessment merely provides an estimate of a person's chronological age. In general, there currently exists no age assessment method that can provide accurate results for the chronological age of a person [35]. In practice, the technique is therefore often used to derive a range of possible ages rather than a precise age.

2.4 Machine Learning

Machine learning is a field that focuses on the development of algorithms and models that can perform tasks "without being explicitly programmed" [69] and instead "learn from experience" [70]. It is a subfield of artificial intelligence (AI), which is an umbrella term for everything related to intelligent machines that can perceive their environment, process information, appropriately respond, and solve problems in a manner similar to humans. Moreover, machine learning can be broadly divided into three paradigms: supervised learning, unsupervised learning, and reinforcement learning. The models used in the publications that form this dissertation fall into the category of supervised machine learning. They are trained to map the input data to a known label, i.e., the desired outcome, which is provided during training. Afterward, a successfully trained model can be applied to make predictions on unseen, unlabelled data, during the so-called inference. For instance, a model may be trained to categorize images by mapping an image to a specific class like cat, dog, car, house, or person.

There are various classical machine learning algorithms and deep learning models available for solving different problems, such as regression, classification, and complex computer vision tasks. Classical machine learning algorithms comprise a range of algorithms and techniques that rely on engineered features and statistical methods to learn patterns from the data. Deep learning models are composed of multiple processing layers that automatically learn feature representations of the input data. Deep learning techniques are a subfield inside the broad field of machine learning. Figure 2.16 illustrates the differences between (non-machine learning) rule-based systems that have to be explicitly programmed, classical machine learning algorithms, and deep learning models.

This section briefly introduces the Random Forest machine learning algorithm, the feature engineering technique Radiomics, and deep learning. Additionally, the role of machine learning in radiology is discussed. In-depth explanations and mathematical details of the most important machine learning methods and concepts can be found in the books *Pattern Recognition and Machine Learning* by Bishop [71] and *An Introduction to Statistical Learning* by James [72]. Deep learning, in particular, is also described in the books *Neural Networks and Deep Learning* by Nielsen [25] and *Deep Learning* by Goodfellow [73]. General commentary on the past, present, and future of machine learning and its role in medicine can be found in various publications, among others in journal articles from LeCun et al. [74], Chartrand et al. [24] and Haug et al. [5].

2.4.1 Random Forest

A random forest is an ensemble of decision trees built with bootstrap aggregating [75] and random feature selection [76, 77]. The algorithm was initially proposed by

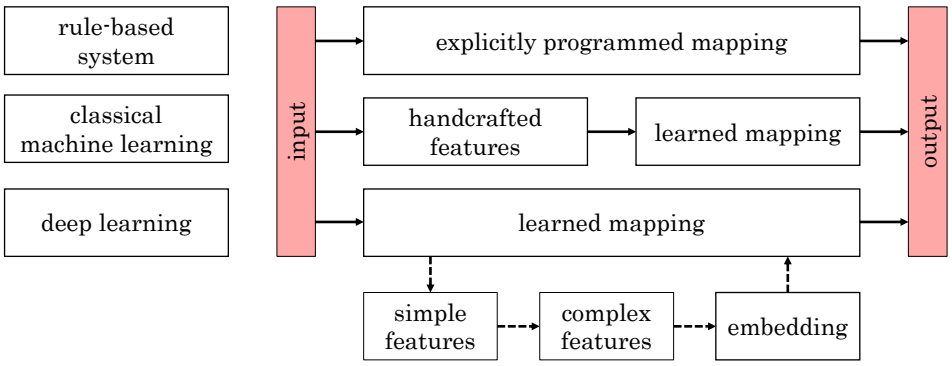

Figure 2.16: Schematic illustration of the differences between rule-based systems, classical machine learning, and deep learning. In rule-based systems, the mapping from input to output is explicitly programmed. Classical machine learning uses a set of given features that are related to the input and learns a mapping between these features and the output. Deep learning directly learns a mapping from input to output, features are learned implicitly during model training. Adapted from [24].

Breiman in 2001 [22] and remains a popular choice for classical machine learning. Each decision tree in the random forest is trained using a special random subset of the entire training data, a so-called bootstrap resample. The nodes inside the tree which recursively split the data are constructed from randomly chosen feature subsets. This twofold randomness serves to grow decision trees that are independent to the largest extent possible. In combination with ensembling, the generalization error of the random forest converges "almost surely to a limit as the number of trees becomes large" [22]. Additionally, random forests provide internal error estimates and allow for feature importance analysis.

Decision trees Decision trees are hierarchical models that make predictions for feature vectors based on a series of binary decisions (Figure 2.17). They are tree-like structures consisting of nodes, branches, and leaves. At each node, a specific feature of the input vector is evaluated. Depending on whether the value of that feature is higher or lower than a previously set threshold, the model progresses down the right or left branch, respectively. In each branch, there is either a new node, i.e. a new binary decision, or a leaf that holds the value of the prediction, i.e. the outcome.

In supervised machine learning, decision trees can be constructed from examples through recursive partitioning based on features, enabling predictions in classification and regression tasks. Given a set of inputs $X = \{\vec{x}_i\}$ with corresponding labels $Y = \{y_i\}$, where each input is a vector of features $\vec{x} = (f_1, ..., f_n)^T$, a decision tree is built by recursively adding nodes to the tree that split the initial set of examples X into subsets X_{left} and X_{right}. These splits are based on a decision criterion $c = (f, t)$,

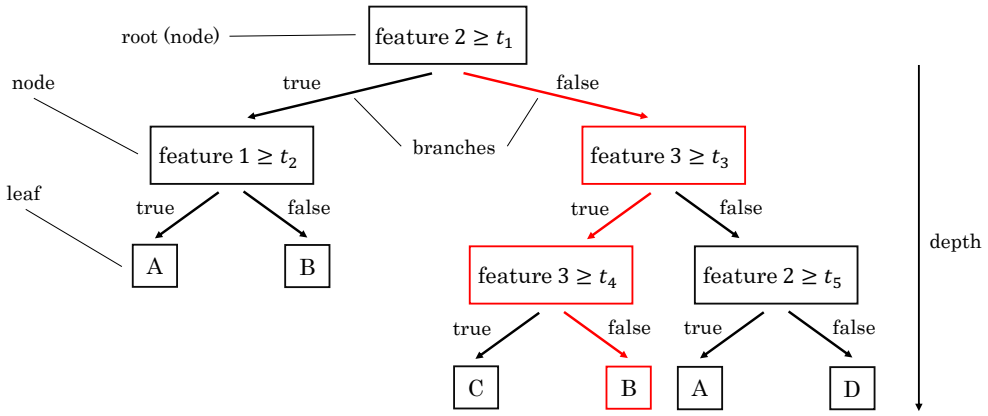

Figure 2.17: Example decision tree for multi-class classification. The model predicts four outcomes, A, B, C, or D, for input feature vectors of length 3, $\vec{x} = (f_1, f_2, f_3)^T$, by comparing individual features to one of five thresholds (t_1 to t_5). The red path illustrates the model's decision-making process (prediction = class B) for inputs with $f_2 < t_1$, $f_3 \geq t_3$, and $f_3 < t_4$.

which consists of a feature f and a threshold t:

$$X_{\text{left}} = \{X \mid f \leq t\} \tag{2.18}$$

$$X_{\text{right}} = \{X \mid f > t\} \quad . \tag{2.19}$$

The starting node of the decision tree is referred to as the root (or root node), while the final subsets are known as leaves (or leaf nodes). Each split aims to minimize the impurity measure H, with the objective of creating subsets that are as pure as possible, i.e. they share the same label:

$$H = \frac{n_{\text{left}}}{n_{\text{left}} + n_{\text{right}}} F(X_{\text{left}}) + \frac{n_{\text{right}}}{n_{\text{left}} + n_{\text{right}}} F(X_{\text{right}}) \quad . \tag{2.20}$$

This process is also referred to as information gain. Ideally, the application of more splits leads to purer subsets.

In classification tasks, the Gini index G is a commonly used impurity measure. It is calculated as the sum of probabilities p for finding a sample $\vec{x}_i$ with the label $y_i = j$ in the subset X':

$$G = \sum_j p_j(1 - p_j) \quad \text{with} \quad p_j = \frac{1}{|X'|} \sum_i I(y_i = j) \quad . \tag{2.21}$$

For regression tasks, the mean squared error can be used to quantify the similarity of labels within a subset:

$$\text{MSE} = \frac{1}{|X'|} \sum_i (y_i - \overline{Y}')^2 \quad . \tag{2.22}$$

The recursive splitting process continues until a specific stopping criterion is met or only one sample remains. A popular stopping criterion is tree depth, i.e. the number of consecutive nodes in a branch. The resulting subgroups from the recursive splits form the leaves of the decision tree.

Random forest algorithm A random forest is trained by constructing an ensemble of decision trees using a specific algorithm. An important hyperparameter that needs to be defined before training is the total number of trees in the random forest, also known as the number of estimators. Given a set of inputs $X = \{\vec{x}_i\}$ with corresponding labels $Y = \{y_i\}$, the construction of each decision tree starts by sampling with replacement from X to create a bootstrap dataset X_{BS}. Typically, about one-third of the instances in X are left out in each bootstrap dataset [22]. Each tree is built using only the samples from its unique bootstrap resample X'. This concept of using bootstrap resamples to train models which are later aggregated is called bagging [75]. The tree construction process is the same as described above for ordinary decision trees, except that the data is recursively split based on a random subset of features for impurity minimization [22]. The random feature subset is drawn at each node.

Random forest predictions are based on the individual predictions made by each decision tree. In classification problems, a majority vote system is applied, where the class with the most votes becomes the final prediction. For regression problems, the final prediction is the average of the individual predictions made by the trees.

Out-of-bag error The data left out of a bootstrap resample is called out-of-bag (OOB) data

$$X_{\mathrm{OOB}} = X \setminus X_{\mathrm{BS}} \tag{2.23}$$

and can be used to acquire accurate estimates of important random forest quantities, e.g. the prediction performance on unseen data. Because each tree has never seen its out-of-bag data X_{OOB} during training, it can be evaluated on these left-out samples. The prediction error of a tree on the corresponding out-of-bag data is called out-of-bag error. The aggregated out-of-bag errors from all trees provide a nearly optimal generalization error of the random forest [78].

Feature importance In the context of medical machine learning applications, understanding the influence of specific features on a model's decision-making process and prediction performance is crucial. The random forest algorithm allows estimating the importance of individual features in multiple ways.

One approach, introduced by Breiman in his original publication [22], is to measure feature importance through feature permutation and the use of out-of-bag data. The idea is to randomly permute the values of a particular feature in the out-of-bag data and to evaluate model performance for these modified subsets of unseen data that

include the noisy feature. By comparing the performance of the permuted out-of-bag datasets with the evaluation results of the original unmodified out-of-bag data, feature importance can be assessed. A larger difference in performance due to feature value permutation indicates higher importance of that particular feature.

Feature importance can also be analyzed using the mean decrease in impurity (MDI) [79]. The average MDI of a specific feature is calculated across the decision trees in the random forest. A higher MDI indicates a greater predictive power of that feature.

Another common strategy for feature importance estimation is to measure the minimal depth [80] at which a particular feature is used to split the data inside the decision trees, on average. Features that are closer to the initial root node split larger portions of the initial dataset more effectively and can be considered more important.

2.4.2 Radiomics

Radiomics refers to the process of extracting and analyzing a large number of quantitative image features (typically exceeding 200) from medical images [23]. This process transforms the images into mineable high-dimensional data that can be effectively analyzed using statistical and machine learning methods to build descriptive and predictive models [28]. The underlying hypothesis of radiomics is, that the advanced analysis of medical images can reveal additional information that is otherwise not used [23]. Specifically, radiomics aims to quantify various phenotypic characteristics, which could potentially provide insights into biological properties such as intra- and inter-tumor heterogeneities [26].

The radiomics workflow (Figure 2.18) can be divided into four steps: imaging, segmentation, feature extraction, and analysis. In the first step, medical images like CT scans, MRI scans, positron emission tomography (PET) scans, or radiographs are acquired. The second step involves identifying regions of interest (ROIs) within the images, which are then segmented through automated segmentation methods or by medical experts. Typical ROIs found in radiomics analysis are lesions, tumors, or abnormalities [81].

In the third step, quantitative imaging features characterizing shape, first-order gray-level histogram statistics, and texture are extracted from these regions. Shape-based features describe the geometric properties of the ROI, such as the diameter along a specific axis, total area, or sphericity. First-order statistics include properties of the intensity distribution inside the ROI, e.g. minimum, maximum, mean, or standard deviation of pixel or voxel gray values. Texture features capture patterns and spatial relationships within the image, such as tissue heterogeneity or coarseness. Often, texture features are computed from matrices like the gray-level co-occurrence matrix (GLCM) or gray-level run-length matrix (GLRLM). Additionally, image filters, such

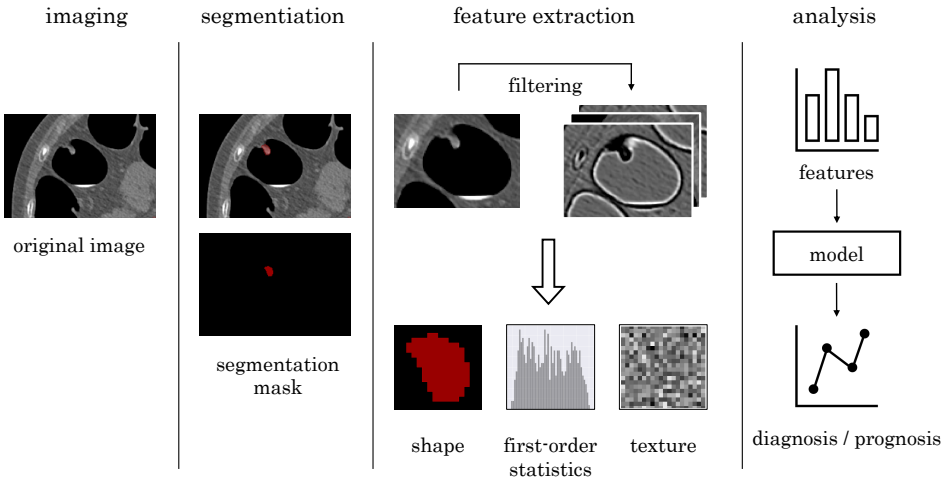

Figure 2.18: The four steps of a typical radiomics workflow (from left to right): imaging, segmentation, feature extraction, and analysis. The image used in this example is a slice from a CT colonography scan and shows a colorectal polyp.

as Laplacian of Gaussian (LoG) filters or wavelet filters, can be applied before feature extraction to highlight particular image properties and capture as much potentially valuable information as possible.

In the fourth step, statistical models and machine learning algorithms are applied to analyze the features or fitted to predict clinical endpoints based on the features. The goal is to uncover relationships between radiomic features and clinical outcomes, such as disease diagnosis, tumor staging, treatment response, or patient survival. Radiomics applications have gained significant interest in recent years in the fields of radiology and radiation oncology [81] and their potential has been demonstrated for multiple tumor types and data extracted from different imaging modalities [26]. To enhance standardization and improve reproducibility in these applications, open-source computation platforms like PyRadiomics [26] and guidelines such as the Image Biomarker Standardization Initiative (IBSI) [82] have emerged.

2.4.3 Deep Learning

Deep learning is a category of machine learning methods, that use multiple layers of non-linear transformations to process data. These layers are fitted to encode a representation of a given set of data. Starting from the input layer, the representations gradually become more abstract, allowing the model to learn complex functions. The core principle of deep learning is that these representations are not designed by hand using domain knowledge, but instead learned entirely from the data using a general-purpose learning procedure called backpropagation [25, 74]. This is particularly useful for image analysis, because deep learning models, especially convolutional

neural networks (CNNs), can directly handle multidimensional pixel arrays [24]. In contrast, classical techniques like a random forest require transforming image data into a feature vector, breaking up the spatial relationship between neighboring pixels, or radiomic features (Figure 2.16).

Deep learning enabled breakthroughs in image, video, text, speech, and audio processing [25]. It is a highly active and dynamic research field with both well-established methods like fully connected networks or CNNs, as well as emerging methods like transformers [31] or latent diffusion models [32], which excel at various tasks such as language processing or image synthesis. However, these latest approaches typically require much larger training datasets and more computing resources compared to fully connected networks and CNNs.

Neural networks Neural networks are a fundamental method of deep learning. They consist of interconnected layers of artificial neurons that use nonlinear transformations to map the information from input to output [71]. The transformations are adaptive, such that they can be learned by the network during training.

Mathematically, neural networks are universal function approximators, and "any continuous function can be uniformly approximated by a continuous neural network having only one internal, hidden layer and with an arbitrary continuous sigmoidal nonlinearity" [83]. Neural networks emerged from attempts to find a mathematical model for the information processing of the nervous system of biological systems. The nonlinearity represents the all or none character of biological neurons that "at any instant have some threshold, which excitation must exceed to initiate an impulse" [84].

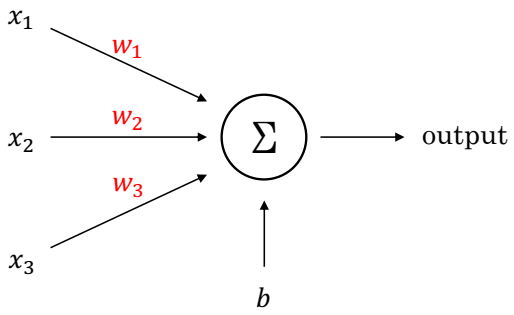

Figure 2.19: Artificial neuron with inputs x_i, weights w_i, and bias b

At the heart of neural networks are artificial neurons consisting of inputs x, weights w, and a bias b (Figure 2.19). The neuron's output a is the sum of the weighted inputs and bias:

$$a = \sum_i w_i x_i + b \quad . \tag{2.24}$$

In a basic neural network, neurons are stacked into layers that take the information from the previous layer $l-1$, transform it, and pass it on to the next layer l (Figure 2.20). Because the information flow is only in one direction, this group of networks is called feed-forward networks. In a feed-forward network, the value of a neuron x_j in layer l is calculated using a feed-forward function g and depends on all neurons x_i from the previous layer $l-1$, the weights w_{ji} connecting x_i and x_j, the bias b_j and the non-linear activation function σ:

$$\vec{x}^l = g(\vec{x}^{l-1}) = \sigma(W^l \vec{x}^{l-1} + b^l) \tag{2.25}$$

$$x_j^l = \sigma\left(\sum_i w_{ji}^l x_i^{l-1} + b_j^l\right) \quad . \tag{2.26}$$

The network output $f(\vec{x}) = \bar{y}$ is calculated by consecutively applying the feed-forward operation to each layer, using the respective weights and biases:

$$f(\vec{x}) = g(\vec{x}^l) = g(...g(g(\vec{x}^0))) = \bar{y} \tag{2.27}$$

Because each neuron of the previous layer influences every neuron in the next layer, these networks are referred to as fully connected feed-forward networks (or short: fully connected networks).

The non-linear activation function maps the neuron's output into a specific range, e.g. $[0, 1]$ or $[-1, 1]$. This way, some inputs yield a low activation ($\sigma(a) = 0 \vee -1$), while others cause a high activation ($\sigma(a) = 1$), which is similar to a biological neuron that only fires if the input signal is above a certain threshold. Common activation functions are the sigmoid function (Equation 2.28) and the rectified linear unit (ReLU) function (Equation 2.29):

$$\text{Sigmoid} \quad \sigma(x) = \frac{1}{1 + \mathrm{e}^{-x}} \quad , \tag{2.28}$$

$$\text{ReLU} \quad \sigma(x) = \max(0, x) \quad . \tag{2.29}$$

The input layer typically matches the representation of the input data, while the output layer contains the processed information in the desired form for the respective task. Between the input layer and output layer are the so-called hidden layers (Figure 2.20).

Neural network training Neural networks are trained using a general-purpose learning procedure called error backpropagation, originally proposed by Rumelhart et al. in 1986 [6]. The training is based on the assumption that a change in any weight or bias at some layer in the network can cause a change in the output. Given a set of inputs $X = x_i$ with corresponding labels $Y = y_i$, the goal of the training is to find a set of weights and biases through repeated adjustment such that the network f approximates $f(x_i) = y_i$ as good as possible for all inputs. To this end,

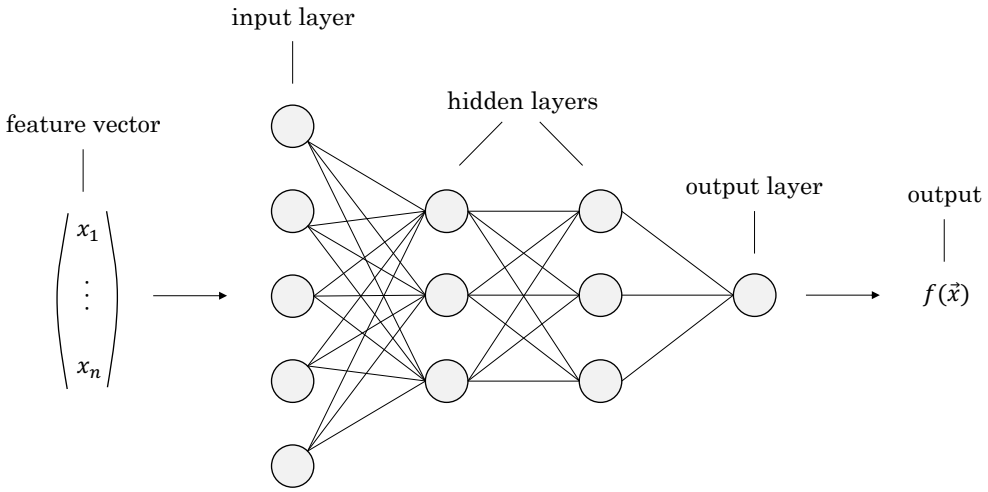

Figure 2.20: Schematic illustration of a fully connected neural network with an input layer, two hidden layers, and an output layer processing a feature vector $\vec{x}$ and returning an output $f(\vec{x})$.

"the procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector" [6]. Mathematically, error backpropagation is a computationally efficient method to compute the derivative of a cost function with respect to all weights and biases in a neural network [71]. The adjustment of weights and biases is performed by an optimization algorithm, based on the cost function derivative calculated with error backpropagation.

The cost function C, also called loss function, is an important element of the training process and is selected based on many criteria, including the task that should be learned. For regression problems, a popular cost function is the mean squared error of all n predictions $f(x_i)$:

$$C = \frac{1}{2n} \sum_{i}^{n} (f(x_i) - y_i)^2 \quad . \tag{2.30}$$

In classification problems, a common cost function is cross-entropy:

$$C = -\frac{1}{n} \sum_{i}^{n} [y_i \ln(f(x_i)) + (1 - y_i) \ln(1 - f(x_i))] \quad . \tag{2.31}$$

Optimization algorithms try to find and apply the optimal change to weights and biases based on the gradient information provided by error backpropagation [85]. Popular optimization algorithms are stochastic gradient descent (SGD) [86, 87] and adaptive moment estimation (ADAM) [88]. SGD iteratively updates weights and

biases by applying a small change in the direction of the negative gradient of the cost function. ADAM uses adaptive first and second-order moments of the gradient estimates in order to update weights and biases.

Convolutional neural networks Convolutional neural networks (CNNs) are another central method of deep learning, which has been widely successful in image- and signal-processing [74]. CNNs address some of the shortcomings of neural networks, e.g. that fully connected networks do not take the spatial structure of images into account. Images have to be transformed into a vector to be used as input in fully connected networks, which disrupts the spatial relation of neighboring pixels.

CNNs are based on three core concepts: local receptive fields, shared weights, and pooling [25]. The idea is to use convolutional units with a receptive field of view, also referred to is kernel or filter, that are shifted across the input data. The first network of that kind was the *Neocognitron* developed by Fukushima in 1980 [89] and has been inspired by the visual nervous system of cats. The method was first applied to a real-world problem successfully by LeCun et al. in 1989 [90] who developed a model for handwritten digit recognition. The term *convolutional neural network* was shaped later in 1998, also by LeCun et al. [91].

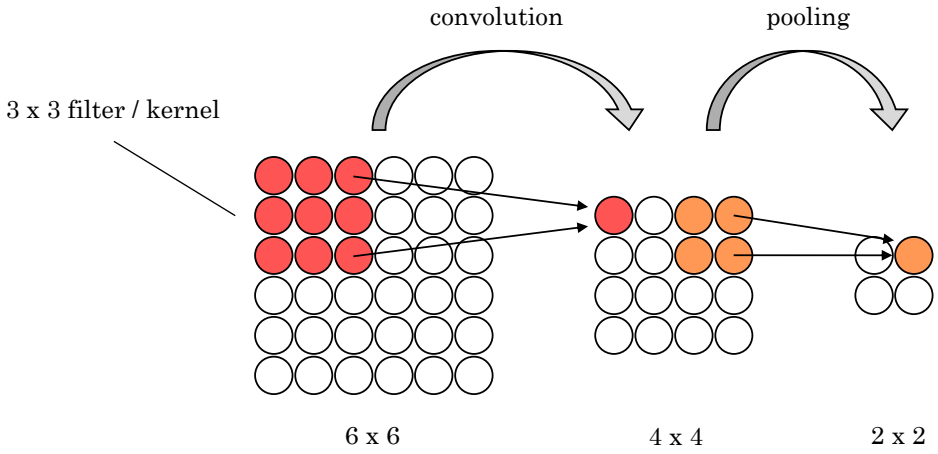

Figure 2.21: Schematic illustration of a convolutional layer followed by a pooling layer typically found in a convolutional neural network. The 3×3 convolution kernel is shifted across the 6×6 input, resulting in a 4×4 feature map output. Next, a pooling operation with a 2×2 field of view simplifies the feature map from the previous layer to a 2×2 feature map.

Given a single-channel two-dimensional image input, e.g. a grayscale radiograph, each pixel is represented by a neuron in the two-dimensional input layer of the CNN. The information from the input layer is mapped to the next hidden layer using the

following convolutional operation

$$x_{ij}^{l} = \sigma\left(\sum_{m}\sum_{n} w_{mn}^{l} x_{i+m,j+n}^{l-1} + b\right) \quad , \tag{2.32}$$

where x_{ij}^{l} is a neuron in layer l, x_{mn} a weight of the kernel with a receptive field of size $m \times n$, b the bias and σ the activation function. Because the same set of weights w_{mn} is shifted across the image (Figure 2.21), the neurons in the hidden layer detect the same feature, extracted from different positions in the image. Therefore, Equation 2.32 is also referred to as a feature map. This complements the translation invariance of images because a specific object in the image will result in the same features, no matter where it is located. On a high level, this allows the same kernel of a CNN to detect a specific structure, e.g. a colorectal polyp, across the entire image, e.g. on the right or left side of the intestinal wall. Additionally, this weight-sharing greatly reduces the number of learnable parameters compared to fully connected networks [25]. This allows either for slimmer and faster networks that offer the same performance or to build deeper and more complex networks that run at the same computational cost.

The sets of shared weights and biases in Equation 2.32 are called kernel or filter. Typically, a CNN has multiple kernels in every layer in order to extract a variety of features, e.g. one kernel may detect vertical edges, while another kernel might detect horizontal ones. With multiple kernels present, the feature mapping from Equation 2.32 becomes more complex (Figure 2.22). First, each kernel calculates its own feature map in the next hidden layer, called a channel. Second, a kernel also calculates the feature map based on all feature maps (channels) from the previous layer.

Convolutional layers are commonly followed by pooling layers, which simplify the information in each feature map. For instance, a common pooling operation is max pooling, where only the neuron with the highest activation in a defined field of view, e.g. 2 x 2 pixels, is carried on to the next layer (Figure 2.22). Intuitively, max-pooling is a way to check whether a feature has been detected or not, without caring about where exactly it has been detected [25]. Other pooling functions such as average pooling are used as well.

Finally, the last layers in CNNs are typically fully connected layers. To this end, the feature maps are at some point flattened and transformed into a feature vector (Figure 2.22). The fully connected layers map the abstract image features from the last layer of feature maps to the output.

Object detection Object detection networks form a particular group of deep learning models that can detect and draw bounding boxes around objects of a particular class in an image. Early networks like R-CNN [92], Fast R-CNN [93], and Faster R-CNN [94] followed a two-stage approach. In the first stage, the network collects

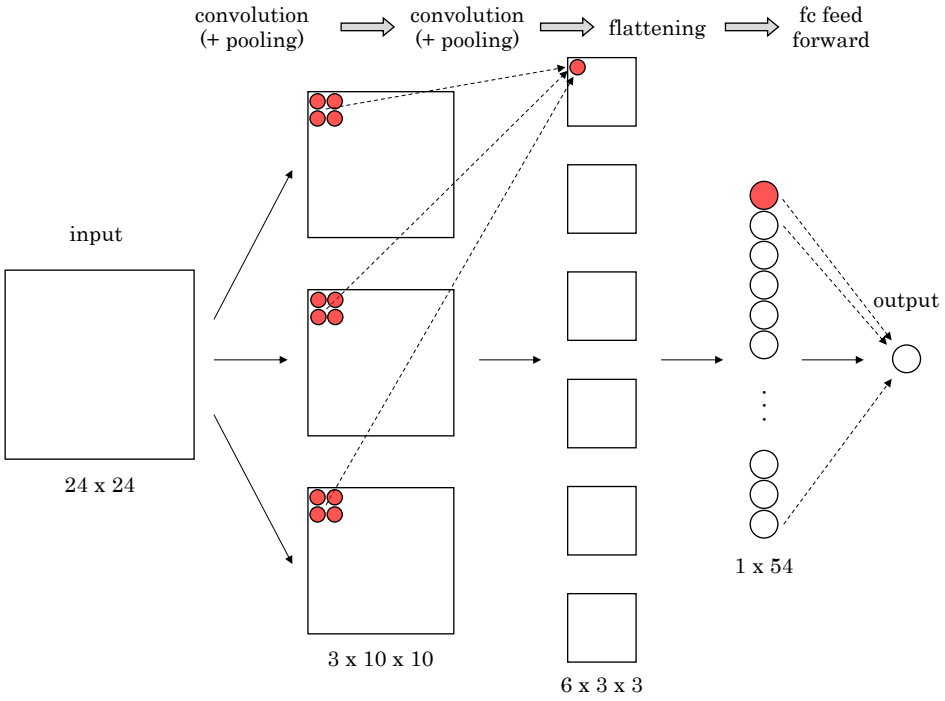

Figure 2.22: Schematic illustration of the information flow in a small example convolutional neural network. First, three convolution kernels are applied to the input, which results in three feature maps. Next, six convolution kernels process the features extracted in the three feature maps from the previous layer, which results in six feature maps. Afterward, the six feature maps are flattened and processed using a fully connected (fc) feed-forward layer that yields the output.

so-called region proposals using techniques like selective search or dedicated CNNs. The second stage is the actual classification step and is performed with classical machine learning algorithms like support vector machines or CNNs.

Driven by the need for faster object detection for real-time video analysis, single-stage object detection models like YOLO [95] and RetinaNet [96] emerged. The RetinaNet was used for the localization of a characteristic anatomical structure in CT scans and is therefore explained briefly in the followng. It uses a backbone network, e.g. an off-the-shelf ResNet, to compute a so-called feature pyramid [97] of the input image. Each level in the feature pyramid is used for detecting objects at a different scale. Based on these features, two separate subnetworks with a simple design perform the object classification and bounding box regression, respectively (Figure 2.23).

An important part of the original implementation of the RetinaNet is the FocalLoss loss function [96], which addresses the problem of heavy class imbalance between foreground and background objects. For each detection, the RetinaNet returns three

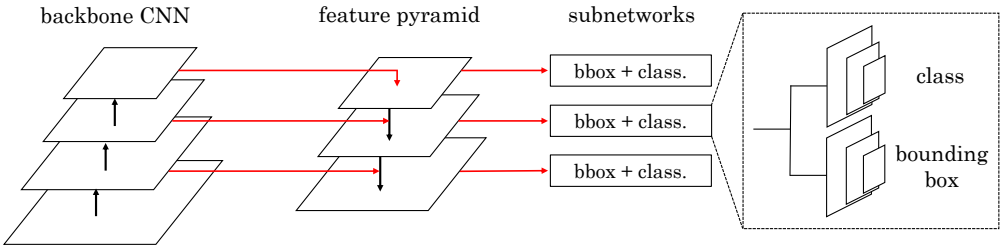

Figure 2.23: Schematic illustration of the RetinaNet object detection network. It uses feed-forward backbone CNN to generate a multi-scale convolutional feature pyramid. The information from the feature pyramid is fed into two subnetworks, one for predicting the class of the object, and one for predicting the anchors of a bounding box. Adapted from [97].

outputs. First, a bounding box prediction that locates the detected object. Second, a class prediction that classifies the detected object. And third, a classification score, between 0.0 and 1.0, that quantifies the confidence of the network in the predicted detection. A classification score above a selected threshold, e.g. ≥ 0.05, is considered a positive detection. The detection is true positive, if the intersection over union (IoU) (Equation 2.33) for the areas of the predicted bounding box A and the ground-truth bounding box B is above a second selected threshold, e.g. ≥ 0.5 and the predicted class is the ground-truth class.

$$IoU = \frac{A \cap B}{A \cup B} \tag{2.33}$$

A popular metric for evaluating object detection performance is average precision since it was applied for the PASCAL Visual Object Classes (VOC) Challenge in 2007 [98, 99]. AP is calculated as the area under the precision-recall curve from all positive and negative network detections, ranked according to classification score in descending order, where the precision p is set to the maximum precision obtained for any recall $r' \geq r$ [98]:

$$AP = \sum_n (r_{n+1} - r_n) \times p_{\text{interp}}(r_{n+1}) \tag{2.34}$$

$$p_{\text{interp}}(r_n) = \max p(r'), \quad r' : r' \geq r_n \tag{2.35}$$

2.4.4 Machine Learning in Radiology

Machine learning has already demonstrated successful applications in certain repetitive medical tasks in the 1990s [5]. For instance, neural networks were used to automate the reading of electrocardiograms [2], counting and classification of white

blood cells [4], and analysis of retinal photographs [1] or skin lesions [3]. Even though early models were simple compared to today's standards, the model performance was already sufficient enough to add clinical value, considering the need for rapid interpretation of an increasing volume of patient data being collected [5]. Today, machine learning is particularly thriving in the fields of radiology and radiation therapy within the medical domain [10, 24, 100, 101].

Radiology departments operate picture archiving and communication systems (PACS) and radiology information systems (RIS) that are used hospital-wide and store vast amounts of image and patient data. These rich databases provide an excellent foundation for the development of machine learning models, as they rely on large datasets for successful training. Consequently, machine learning with medical images is a highly active and dynamic research field [100, 102–104]. Additionally, the increasing number of machine learning-enabled medical devices in radiology approved by the U.S. Food & Drug Administration [9] shows the transition of more and more machine learning research into clinical practice (Figure 2.24).

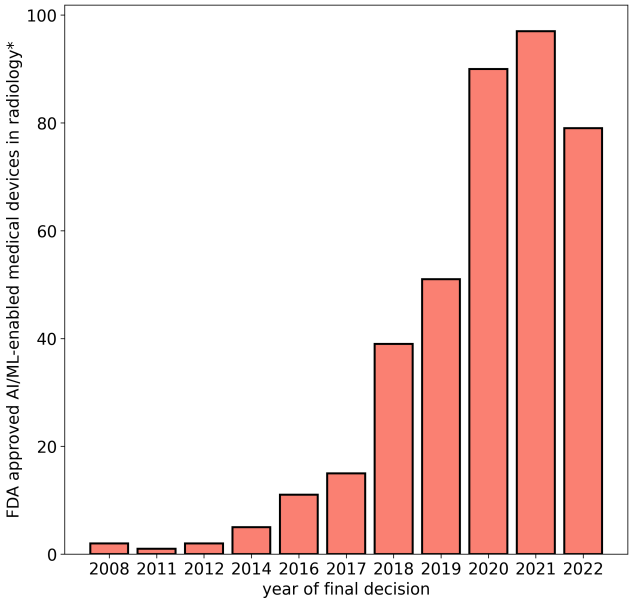

Figure 2.24: FDA-approved artificial intelligence (AI)/machine learning (machine learning)-enabled medical devices marketed in the United States in radiology (status: October 5, 2022). The numbers are based on publicly available resources (*) and other publicly available materials published by specific manufacturers. Data from [9].

Problem categories Machine learning in radiology commonly addresses the following problem categories: image classification, semantic segmentation, instance segmentation, and object detection [10, 24, 105]. Image classification is the task of

assigning one or multiple classes (labels) to an image, e.g. a specific lesion category or medical condition. In semantic segmentation, an image is partitioned into multiple segments or regions, e.g. into different organs, or tumors and healthy tissue. Instance segmentation is similar, but the problem is formulated as distinguishing different instances of a particular class within an image. In object detection, the task is to detect one or multiple objects of a particular class in an image, e.g. by drawing bounding boxes around all lung nodules in a thoracic CT scan.

Challenges While machine learning is a powerful tool in radiology today, there are still a number of challenges that need to be addressed in order to make applications as safe, fair, and accurate as possible.

One problem is the availability of annotated and structured data. Supervised machine learning, particularly deep learning, relies on large amounts of data with annotations in a machine-readable format [74]. However, only a small fraction of medical images stored in hospital archives worldwide are appropriately annotated for most research purposes. In some cases, the required information for data labeling exists, but is not available in a machine-readable form, because it may have been recorded in writing inside a radiology report. Other times the annotations don't exist at all, in particular, labels for segmentation and object detection tasks are rare. Data annotation for machine learning in radiology typically requires human experts, which makes it a time-consuming and expensive process that poses an enormous bottleneck for machine learning development [41, 103]. Classical machine learning methods can be less data-hungry than deep learning and, when combined with feature extraction techniques like Radiomics, enable automated image analysis also for smaller datasets.

Another issue concerns data quality. Especially in the sensitive medical environment, training data for machine learning models should be diverse, unbiased, and accurately labeled [103]. For instance, images from a single hospital can be insufficient to train a model, because it may be biased toward the sampled population and acquisition scheme [10].

Furthermore, there are technical and legal barriers to sharing medical image data due to privacy concerns. The sensitive nature of medical information requires strict privacy protocols, which can obstruct the sharing of data and collaboration of researchers across institutions.

Finally, the medical community expects the same level of certainty for a machine learning tool as for a drug or any other classical type of intervention method [5]. However, the lack of clear standards for describing and evaluating AI and machine-learning applications makes it difficult to identify reliable tools and establish trust in machine learning amongst clinicians.

Outlook The current consensus is that machine learning will be a valuable assistant for clinicians in radiology, and medicine in general [5]. Machine learning will not

replace doctors, but help them to do their jobs better and hopefully free up valuable time for human doctor-patient interactions. The patient–doctor relationship will remain the cornerstone of patient care, and machine learning has the potential to enrich that relationship [103].

Today's research is also going to shape the type of machine learning models deployed in clinical practice in the future. For instance, most available FDA-approved machine learning tools have approval for very specific and narrow tasks. However, some experts envision the future of machine learning in the development of general multi-purpose models, called foundation models [106]. These are large models trained on vast and diverse datasets with a wide range of input and output formats, typically in an unsupervised fashion [107]. The idea is to create highly flexible models with domain knowledge that can be fine-tuned for a wide range of downstream tasks.

Going forward, controlled studies measuring practical clinical endpoints are necessary to better understand the clinical value and quantify the impact of machine learning in radiology [10].

3 | Contributions to original publications

This chapter summarizes my contributions to the three original publications and the complementing publication upon which this cumulative dissertation is based.

3.1 Contributions to Original Publication I

The first publication (chapter 4) entitled *Machine Learning-based Differentiation of Benign and Premalignant Colorectal Polyps Detected with CT Colonography in an Asymptomatic Screening Population: A Proof-of-Concept Study* was conceptualized in cooperation with all co-authors.

My contributions to this publication involved data curation, radiomic feature extraction, machine learning model training, validation and testing, general result analysis, radiomic feature importance analysis, result visualization, and writing parts of the original manuscript draft. First, I created a data set for machine learning from computed tomography (CT) colonography scans and tables of clinical parameters. This required carefully matching the CT colonography scans to the histopathologically confirmed colorectal polyp character (benign vs. premalignant) of the respective patient. Next, I calculated radiomic features from the CT colonography scans and the provided manual polyp segmentation masks using the Python package PyRadiomics. Afterward, I trained, validated, and tested a random forest machine learning model for predicting polyp character based on the previously calculated radiomic features. The random forest model was implemented using the Python package scikit-learn. I tested the random forest prediction performance on an external test set. Additionally, I performed a radiomic feature importance analysis to gain insight into the random forest prediction process. Also, I created plots of the methodological workflow and the results. The results were critically discussed with PD Dr. med. Sergio Grosu, Prof. Dr. rer. nat. Michael Ingrisch and PD Dr. med. Philipp Kazmierczak. Together with PD Dr. med. Sergio Grosu., I wrote parts of the original manuscript draft. Finally, I reviewed and edited the manuscript in cooperation with all co-authors.

3.2 Contributions to Original Publication II

The second publication (chapter 5) entitled *Deep learning in CT colonography: differentiating premalignant from benign colorectal polyps* was conceptualized in collaboration with all co-authors.

My contributions to this publication involved data curation, deep learning model training, validation and testing, general result analysis, model interpretation, result visualization, and writing parts of the original manuscript draft. First, I preprocessed the dataset from publication I (chapter 4), such that it could be used for developing a deep learning model. This included cropping the CT colonography scans around the manually segmented colorectal polyps. Cropping was necessary because training a deep learning model on the full CT scans in native resolution was computationally not feasible. Also, a deep learning model for full-size CT scans would have had to be large and complex, which would have made it difficult to train with the few available examples to learn from. Next, I trained, validated, and tested a convolutional neural network (CNN) for predicting polyp character based on the cropped CT colonography scans with and without the manual polyp segmentation masks. The random forest model was implemented using the Python package Keras and the Python machine learning library TensorFlow as a backend. Again, I tested the CNN performance on an external test set. To enable model interpretability, I used the gradient-based visualization technique GradCAM++ to highlight image regions that are potentially important for CNN predictions. Additionally, I created plots of the data, methods, and results. The results were critically discussed with PD Dr. med. Sergio Grosu and Prof. Dr. rer. nat. Michael Ingrisch. I wrote the original manuscript draft with assistance from PD Dr. med. Sergio Grosu. Finally, I reviewed and edited the manuscript in cooperation with all co-authors.

This interdisciplinary study required expertise in the fields of radiology as well as machine learning and data science. Therefore, the first authorship is shared with PD Dr. med. Sergio Grosu, who was a resident in diagnostic radiology at the time of publication.

3.3 Contributions to Original Publication III

The third publication (chapter 6) entitled *Automated localization of the medial clavicular epiphyseal cartilages using an object detection network: a step towards deep learning-based forensic age assessment* was conceptualized in collaboration with all co-authors.

My contributions to this publication involved data curation, deep learning model training, validation and testing, general result analysis, result visualization, and writing the original manuscript draft. First, I defined inclusion and exclusion criteria for the study collective together with PD Dr. med. Bastian Sabel and Dr. rer.

nat. Balthasar Schachtner. Next, I assissted Dr. rer. nat. Balthasar Schachtner in acquiring thoracic CT scans from the hospital's picture archiving and communication system (PACS). I manually drew bounding boxes around a proxy structure for the medial clavicular epiphyseal cartilages in two-dimensional axial slices of thoracic CT scans. Afterward, I trained an instance of the deep learning object detection network RetinaNet to detect this proxy structure where the bounding boxes served as a position label. I validated the model using a dedicated test set. Based on the trained RetinaNet, I developed an algorithm for the automated and unique localization of the medial clavicular epiphyseal proxy structure in full-size thoracic CT scan volumes. The results were critically discussed with PD Dr. med. Bastian Sabel, Prof. Dr. rer. nat. Michael Ingrisch and Dr. rer. nat. Balthasar Schachtner. I wrote the original manuscript draft and created plots of the data, methods, and results. Finally, I reviewed and edited the manuscript in cooperation with all co-authors.

3.4 Contributions to Complementing Publication I (Appendix)

The complementing publication (section A.1) entitled *Radiological age assessment based on clavicle ossification in CT: Enhanced accuracy through deep learning* was conceptualized in collaboration with all co-authors.

My contributions to this publication involved data curation, deep learning model training, validation and testing, general result analysis, result visualization, and writing the original manuscript draft. First, I assissted Dr. rer. nat. Balthasar Schachtner in acquiring additional thoracic CT scans from the hospital's PACS to enrich the dataset of publication III (chapter 6) with additional images. Next, I pre-processed all CT scans using the localization algorithm developed in publication III (chapter 6) and cropped the scans around the medial clavicular epiphyseal cartilages. Afterward, I trained, validated, and tested an ensemble of 20 CNNs for predicting chronological age based on the previously cropped thoracic CT scans. The CNNs were based on the popular ResNet18 architecture for two-dimensional images and adjusted by me to enable the processing of three-dimensional inputs. Additionally, I developed an optimistic human reader performance estimate for classical radiological age assessment by human experts together with Dr. rer. nat. Balthasar Schachtner. Afterward, I compared the age prediction accuracy of the deep learning ensemble with the human reader performance estimate. The results were critically discussed with Dr. rer. nat. Balthasar Schachtner, Prof. Dr. rer. nat. Michael Ingrisch and PD Dr. med. Bastian Sabel. I wrote the original manuscript draft and created plots of the data, methods, and results. Finally, I reviewed and edited the manuscript in cooperation with all co-authors.

4 | Publication I

ORIGINAL RESEARCH • GASTROINTESTINAL IMAGING

Radiology

Machine Learning–based Differentiation of Benign and Premalignant Colorectal Polyps Detected with CT Colonography in an Asymptomatic Screening Population: A Proof-of-Concept Study

Sergio Grosu, MD • Philipp Wesp, MSc • Anno Graser, MD • Stefan Maurus, MD • Christian Schulz, MD • Thomas Knösel, MD • Clemens C. Cyran, MD • Jens Ricke, MD • Michael Ingrisch, PhD • Philipp M. Kazmierczak, MD

From the Department of Radiology (S.G., P.W., S.M., C.C.C., J.R., M.I., P.M.K.), Department of Medicine II (C.S.), and Department of Pathology (T.K.), University Hospital, LMU Munich, Marchioninistr 15, 81377 Munich, Germany; and Radiologie München, Munich, Germany (A.G.). Received May 23, 2020; revision requested June 23; revision received December 10; accepted December 22. **Address correspondence to** S.G. (e-mail: *sergio.grosu@med.uni-muenchen.de*).

Supported by FöFoLe, Faculty of Medicine, Ludwig-Maximilians-Universität München (S.G., P.W.).

Conflicts of interest are listed at the end of this article.

Radiology 2021; 299:326–335 • https://doi.org/10.1148/radiol.2021202363 • Content code: GI

Background: CT colonography does not enable definite differentiation between benign and premalignant colorectal polyps.

Purpose: To perform machine learning–based differentiation of benign and premalignant colorectal polyps detected with CT colonography in an average-risk asymptomatic colorectal cancer screening sample with external validation using radiomics.

Materials and Methods: In this secondary analysis of a prospective trial, colorectal polyps of all size categories and morphologies were manually segmented on CT colonographic images and were classified as benign (hyperplastic polyp or regular mucosa) or premalignant (adenoma) according to the histopathologic reference standard. Quantitative image features characterizing shape (n = 14), gray level histogram statistics (n = 18), and image texture (n = 68) were extracted from segmentations after applying 22 image filters, resulting in 1906 feature-filter combinations. Based on these features, a random forest classification algorithm was trained to predict the individual polyp character. Diagnostic performance was validated in an external test set.

Results: The random forest model was fitted using a training set consisting of 107 colorectal polyps in 63 patients (mean age, 63 years ± 8 [standard deviation]; 40 men) comprising 169 segmentations on CT colonographic images. The external test set included 77 polyps in 59 patients comprising 118 segmentations. Random forest analysis yielded an area under the receiver operating characteristic curve of 0.91 (95% CI: 0.85, 0.96), a sensitivity of 82% (65 of 79) (95% CI: 74%, 91%), and a specificity of 85% (33 of 39) (95% CI: 72%, 95%) in the external test set. In two subgroup analyses of the external test set, the area under the receiver operating characteristic curve was 0.87 in the size category of 6–9 mm and 0.90 in the size category of 10 mm or larger. The most important image feature for decision making (relative importance of 3.7%) was quantifying first-order gray level histogram statistics.

Conclusion: In this proof-of-concept study, machine learning–based image analysis enabled noninvasive differentiation of benign and premalignant colorectal polyps with CT colonography.

Online supplemental material is available for this article.

In industrialized countries, colorectal cancer is among the three most common causes of cancer-related death (1,2). It is assumed that most types of colorectal cancer originate from adenomatous polyps developing over several years (3). Thus, the incidence and mortality of colorectal cancer can be reduced by early detection of precancerous polyps with consecutive resection (4–6). As clinical symptoms are unspecific and often absent, particularly in the early stages, screening procedures such as immunochemical fecal occult blood test and optical colonoscopy (OC) play a major role in colorectal cancer prevention (7,8).

During the past 2 decades, CT colonography emerged as a noninvasive screening method for colorectal cancer. The sensitivity of CT colonography and OC for the detection of colon polyps that are 6 mm or larger (and hence advanced adenoma detection rate) are comparable in asymptomatic screening populations and in patients with symptoms suggestive of colorectal cancer (9–11). Also, CT colonography is effective in visualizing portions of the colon not evaluated by OC in cases of complex anatomic conditions causing failed or incomplete OC and therefore permits robust polyp detection also in the right colon (12). But CT colonography does not enable a definite differentiation between benign and premalignant polyps, crucial for individual risk stratification and therapy guidance. Therefore, polyp size measured in CT colonography data sets is currently used as surrogate indicator of the likelihood of malignancy. Current guidelines recommend OC-guided resection for colorectal polyps that are 6 mm or larger (United States Multi-Society Task Force on Colorectal Cancer, European Society of Gastrointestinal

Abbreviations

AUC = area under the receiver operating characteristic curve, OC = optical colonoscopy

Summary

In this proof-of-concept study, machine learning–assisted CT colonography analysis allowed for differentiation of benign and premalignant colorectal polyps at high diagnostic performance associated with the histopathologic reference standard.

Key Results

- Radiomics-based image analysis enables noninvasive differentiation of benign and premalignant CT colonography–detected colorectal polyps, with an area under the receiver operating characteristic curve of 0.91, a sensitivity of 82%, and a specificity of 85%.
- The area under the receiver operating characteristic curve for the machine learning–based differentiation of benign versus premalignant polyps was 0.87 in the 6–9-mm size category and 0.90 in the 10-mm or larger size category.

Endoscopy, and European Society of Gastrointestinal and Abdominal Radiology) (13,14).

Radiomics describes the process of converting medical image data sets into mineable high-dimensional data by extracting quantitative image features based on intensity, shape, size or volume, and texture. Using machine learning approaches, such as random forests, support vector machines, or neural networks, tumor-specific radiomics features enable comprehensive tumor characterization beyond the visible morphology in radiologic images (15–17). For instance, CT texture features predicted KRAS (Kirsten rat sarcoma viral oncogene homolog) mutation status of manifest colorectal cancer with an area under the receiver operating characteristic curve (AUC) of 0.83 (18).

The aim of this proof-of-concept study was to establish a noninvasive, radiomics-based machine-learning differentiation of benign (ie, hyperplastic polyp or regular mucosa) and premalignant (ie, adenoma) polyps in CT colonography data sets from an asymptomatic, average-risk colorectal cancer screening cohort.

Materials and Methods

Training Set

The present study was approved by the institutional review board, and the requirement for written informed consent was waived. The random forest machine learning model in this retrospective analysis was trained using CT colonography scans from a previously published prospective colorectal cancer screening cohort of an average-risk asymptomatic population older than 50 years undergoing fecal occult blood test, fecal immunochemical stool test, and same-day OC and CT colonography to compare the performance of these screening tests in the detection of advanced colonic neoplasia (9). Participants were enrolled in the previously published colorectal cancer screening cohort only if they had no symptoms of colonic diseases, such as melaenic stools, hematochezia, abdominal pain, relevant changes in stool frequency, diarrhea, inflammatory bowel disease, hereditary colorectal cancer syndromes, positive family history for colorectal cancer (ie, one first-degree relative with colorectal cancer diagnosis before 60 years of age or two first-degree relatives with colorectal cancer diagnosis at any age), prior OC within the past 5 years, body weight greater than 150 kg, and severe cardiovascular or pulmonary disease (9). In the present study, a polyp-enriched data set was created, including only participants with histopathologically characterized polyps (Fig 1).

CT Colonography in the Training Set

Bowel preparation included 4 L of polyethylene glycol solution (KleanPrep; Norgine Pharmaceuticals) and a combination of 20 mg bisacodyl with 30 mL of sodium phosphate (Prepacol; Guerbet Pharma). A volume of 50 mL of the iodinated contrast agent iopamidol (Solutrast 300; BraccoAltana Pharma) was added to the last liter of polyethylene glycol electrolyte solution to tag residual fluid in the colon. Image data sets were acquired with a 64-channel multidetector row scanner (Somatom Sensation 64; Siemens Healthineers) at 0.6-mm collimation. Images were reconstructed at a section thickness of 0.75 mm and a 0.5-mm reconstruction increment using a standard soft-tissue kernel. CT scans in the supine position were acquired at a tube voltage of 120 kVp and tube current–time product reference values of 70 mAs. CT scans in the prone position were acquired at a tube voltage of 120 kVp and tube current–time product reference values of 30 mAs using an online dose modulation technique for automatic tube current adaption (Care Dose 4D; Siemens Healthineers). Dose-length products were recorded as estimates of radiation exposure. Mean radiation dose for CT colonography was 4.5 mSv (0.6). Bowel distention was achieved via manual air or automated carbon dioxide insufflation via a rectal tube and evaluated by a radiologist on the CT scout film of the abdomen. The first set of images was obtained in a 7–9-second breath hold in the supine position, and the second set was obtained after repositioning of the study participant in the prone position. Bowel preparation and the CT colonography protocol were previously described in detail (9). All bowels were adequately distended, cleansed, and tagged to ensure high diagnostic performance of CT colonography.

External Test Set

To estimate generalization performance of our machine learning model, we used external CT colonography data sets from a large North American multicenter CT colonography screening trial made publicly accessible via The Cancer Imaging Archive (19–21). The Cancer Imaging Archive is a large multicenter open-source open-access collection of anonymized medical images of cancer, including radiologic data sets. The CT colonography data sets of the aforementioned multicenter screening trial were acquired with multiple CT scanners from several vendors (Siemens Healthineers, GE Healthcare Systems, Philips Healthcare, Canon Medical Systems) with varying scanning protocols. Only polyps with available histopathologic reports were included. Polyps that could not be unequivocally identified because of poor bowel distension or insufficient

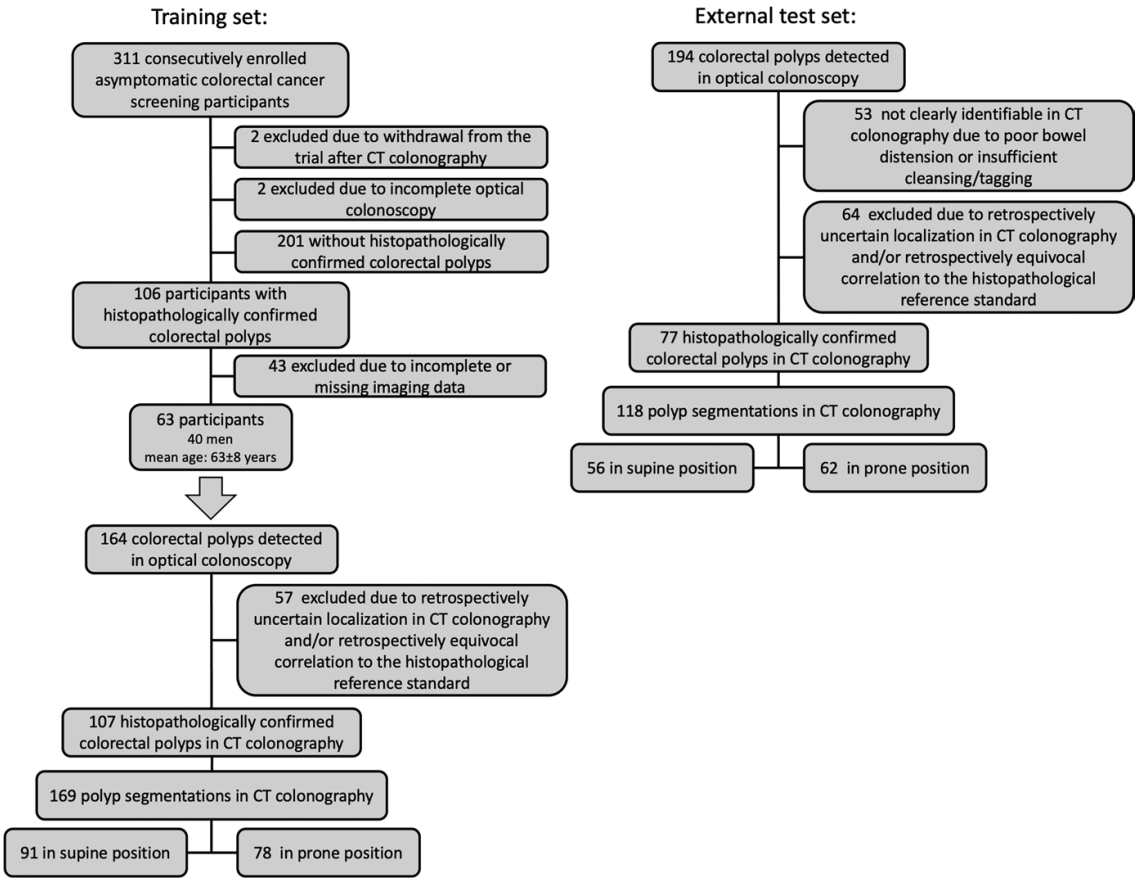

Figure 1: Flow diagram of the training set and external test set.

cleansing or tagging were excluded from the analysis. Polyps with retrospectively equivocal or uncertain correlation to the histopathologic reference standard were also excluded.

Image Segmentation

Colorectal polyps in both data sets were manually segmented for radiomics feature extraction by a board-certified radiologist (P.M.K., 8 years of experience in oncologic whole-body imaging) and two radiology residents (S.G., S.M.; 3 years of experience in oncologic whole-body imaging) in multiplanar two-dimensional CT colonography images (Fig 2). Information on polyp size and colon segment in which polypectomy was performed was provided to all readers. All readers were blinded to the histopathologic polyp class. Polyp segmentations were performed by each reader in equal amounts. Each polyp segmentation was confirmed by the other two readers who did not perform the segmentation. In case of divergent reading results, a consensus reading was performed. To assess intrareader, interreader, and radiomics feature variability, 25% of polyp segmentations from the test set were randomly selected with stratification by class label (benign or premalignant) and were again segmented independently by the same reader (intrareader variability) and a second reader (interreader variability). These second polyp segmentations were performed 5 months after the first segmentations. Combined reading of two-dimensional CT colonography images and virtual fly-through three-dimensional reconstructions was used for exact polyp detection. Manual polyp segmentation was performed with multiplanar two-dimensional CT colonography images. Each colorectal polyp was segmented on supine and prone position scans if clearly identifiable in both positions. If a polyp could be located in only the supine or prone position, it was segmented in that position only. Polyps that could not be securely identified or unequivocally assigned to the corresponding histopathologic report were excluded from analysis. Polyp size categories were 5 mm or smaller, 6–9 mm, and 10 mm or larger. In addition, polyps were morphologically classified as pedunculated, sessile, or flat. The CT colonography workflow of the commercially available dedicated postprocessing software syngo.via version VA30B (Siemens Healthineers) was used for exact polyp detection and localization. The free open-source software Medical Imaging Interaction Toolkit, version 2018.04 (German Cancer Research Center) was used for manual segmentation (22).

Histopathologic Reference Standard

Colorectal polyps were included only if they were unequivocally assignable to the corresponding histopathologic report (Fig 3). A polyp was considered benign if it was diagnosed as a hyperplastic polyp or regular mucosa in the corresponding histopathologic report. A polyp was considered premalignant if it was diagnosed as tubular adenoma, tubulovillous adenoma, or villous adenoma in the corresponding histopathologic report. For study purposes only, a small number of lesions with the histopathologic classification serrated adenoma (four polyp segmentations) or adenocarcinoma (five segmentations) were included in the premalignant group. Correspondingly, two polyp segmentations with the histopathologic classification lipomatous polyp were categorized in the benign group.

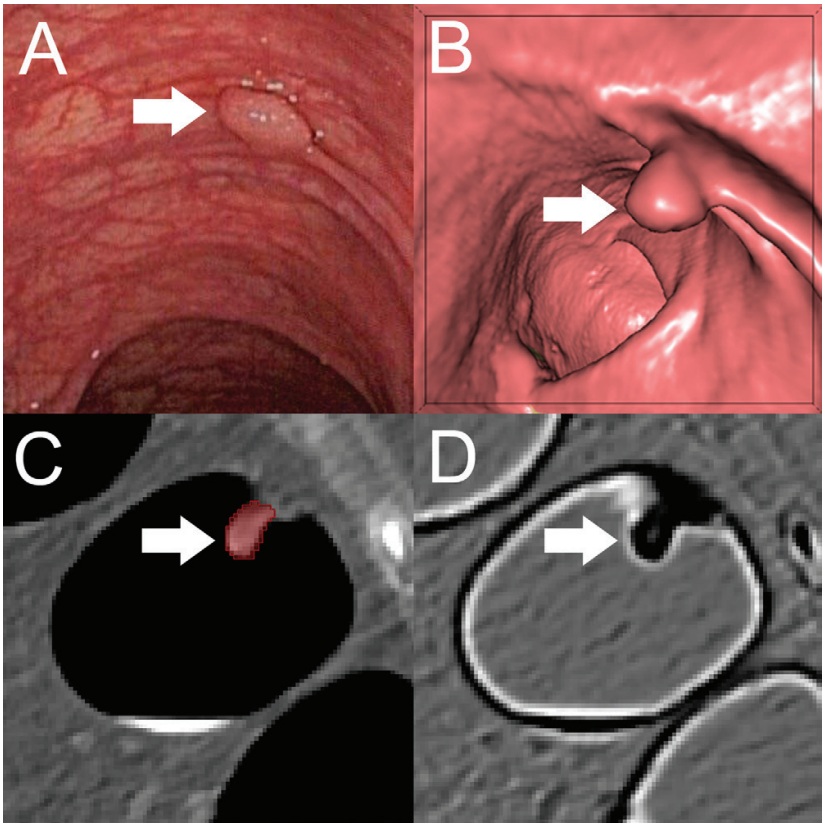

Figure 2: *A*, Optical colonoscopy and, *B–D*, CT colonography of a 9-mm polyp (arrow) in the descending colon of a 78-year-old woman. *B*, Virtual fly-through three-dimensional reconstructions were used for exact polyp localization. *C*, Manual polyp segmentation was performed in multiplanar two-dimensional CT colonography images. *D*, CT colonography images were preprocessed for image feature extraction by application of a dedicated filter.

Feature Extraction

Quantitative image features were extracted from segmented voxels in CT colonography scans using the open-source Python package Pyradiomics (version 2.2.0; Harvard Medical School) (23). Quantitative image features, including gray level histogram statistics (n = 18) and image texture (n = 68), were extracted after applying 22 image filters for each, and 14 characterizing shapes (including size) also were extracted, resulting in 1906 (ie, [18 + 68] · 22 + 14) feature-filter combinations, which will be referred to simply as *features*. Feature extraction is illustrated in Figure 4 as part of the entire radiomics workflow. Lists of all image filters and extracted features are provided in Tables E3 and E4 (online). To facilitate feature reproducibility and comparability of polyps scanned at different centers, all CT colonography scans were preprocessed prior to feature extraction, including a resampling of CT pixel spacing to 0.72 mm along the x-axis and y-axis and 0.5 mm along the z-axis.

Feature Selection

Among the 1906 features previously extracted, features highly correlated with another feature were identified and excluded from the analysis. This feature selection based on pairwise feature correlation was applied to improve the machine learning training process and to enable optimized feature interpretability. For this purpose, we calculated a Pearson correlation matrix for all features extracted from the training set and excluded one feature of any feature pair (1906 × 1905/2 = 1.8 million pairs) with a Pearson correlation coefficient greater than 0.8. The exclusion of features that were too highly correlated reduced the number of features used for the final analysis to 10% (198 of 1906).

Random Forest Training

All 169 polyp segmentations in the training set were class-divided according to the histopathologic reference standard into benign or premalignant. The random forest classifier sklearn.ensemble.RandomForestClassifier (Python Scikit-learn machine learning library, version 0.22 [24]) with 1000 trees (n_estimators = 1000) and otherwise default parameters was trained on the 169 polyp segmentations of the training set using the previously identified set of 198 features to predict polyp class (ie, to differentiate between benign and premalignant colorectal polyps). Each decision tree in the random forest was trained on bootstrap resamples of the entire training data. Inside the random forest tree, binary decisions

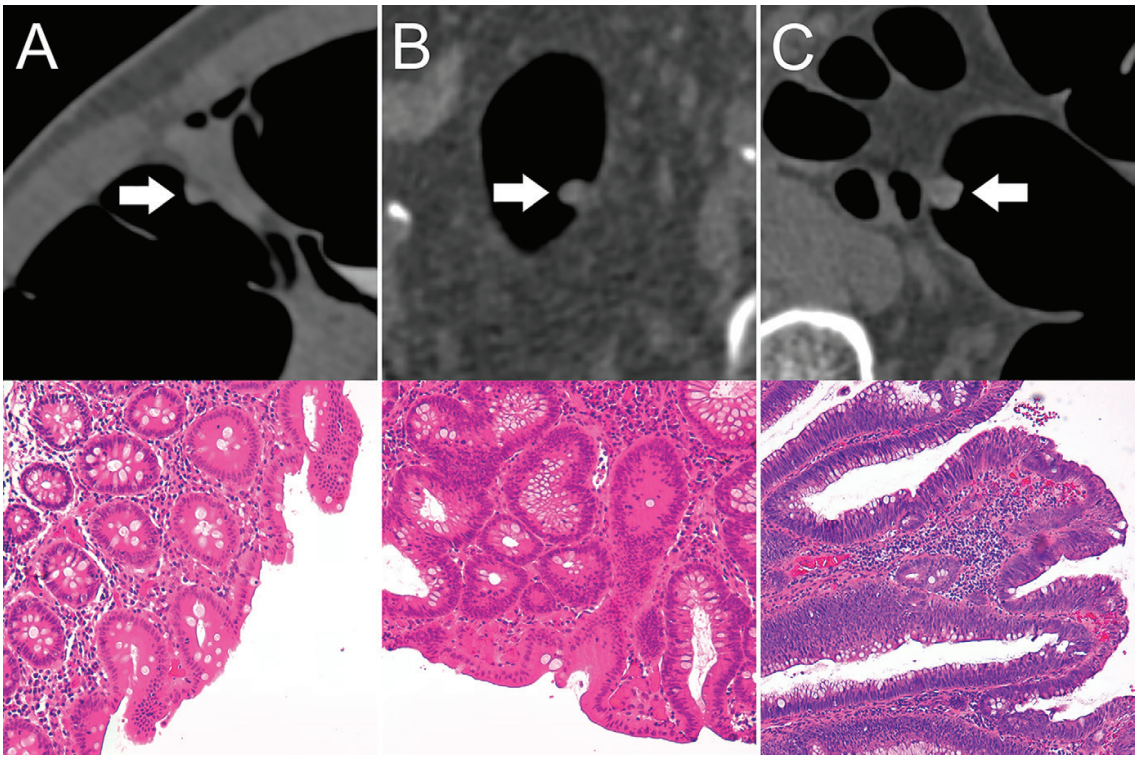

Figure 3: Top: Axial CT colonography images show representative colorectal polyps (arrow) in the training set. Bottom: Corresponding histopathologic work-up. (Hematoxylin-eosin staining; original magnification ×20.) *A*, An 8-mm hyperplastic polyp in the ascending colon of a 54-year-old woman with hyperplastic epithelia. *B*, An 8-mm tubular adenoma in the sigmoid colon of a 68-year-old man with tubular growth pattern and elongated nuclei. *C*, An 11-mm tubulovillous adenoma in the rectum of a 73-year-old man with tubulovillous growth pattern and elongated nuclei.

at individual nodes were learned on randomly chosen feature subsets. This twofold randomness served to grow decision trees that are independent to the largest extent possible, such that "the generalization error converges almost surely to a limit as the number of trees becomes large" (25). The Scikit-learn random forest implementation followed Breiman et al (25) with one minor exception: It combined classifiers by averaging their probabilistic prediction instead of letting each classifier vote for a single class. Compared with other popular machine learning algorithms (eg, AdaBoost), random forests are robust to outliers and noise (25,26). Additionally, random forests provide internal error estimates that can also be used for estimating feature importance.

Statistical Analysis of the Test Set

Diagnostic performance of our machine learning–based polyp classification approach was assessed by evaluating the previously trained random forest on the external test set. Therefore, the 118 polyp segmentations in the external test set were class-divided into benign and premalignant in the same manner as the training set, and the same set of 198 features used for training was extracted per image. Performance was quantified with three metrics: AUC, sensitivity, and specificity. Sensitivity and specificity depended on the classification threshold, a parameter that was used to turn predicted class probabilities, the output of the random forest model for a given input sample, into class predictions (benign vs premalignant). Sensitivity and specificity were evaluated for a default threshold value of 0.5 (model A), a threshold value that maximized the Youden index (J = sensitivity + specificity − 1) (27) (model B), and a threshold that resulted in the highest possible specificity while achieving a sensitivity of at least 85% (model C). For the AUC, sensitivity, and specificity, 95% CIs were calculated from 2000 bootstrap samples of the external test set (28). The random forest analysis is shown in Figure 5.

Statistical Analysis of the Training Set and Model Introspection

Training set samples left out for training a tree, so-called out-of-bag samples, were used to self-evaluate the respective tree and ultimately to form an internal prediction error estimate for the random forest, which is called the out-of-bag error (29). In addition, the Scikit-learn random forest implementation provided an internal estimate of feature importance, namely how much the class prediction (benign vs premalignant) of a trained model depended on a specific feature relative to all other features. We exploited this by evaluating the relative importance

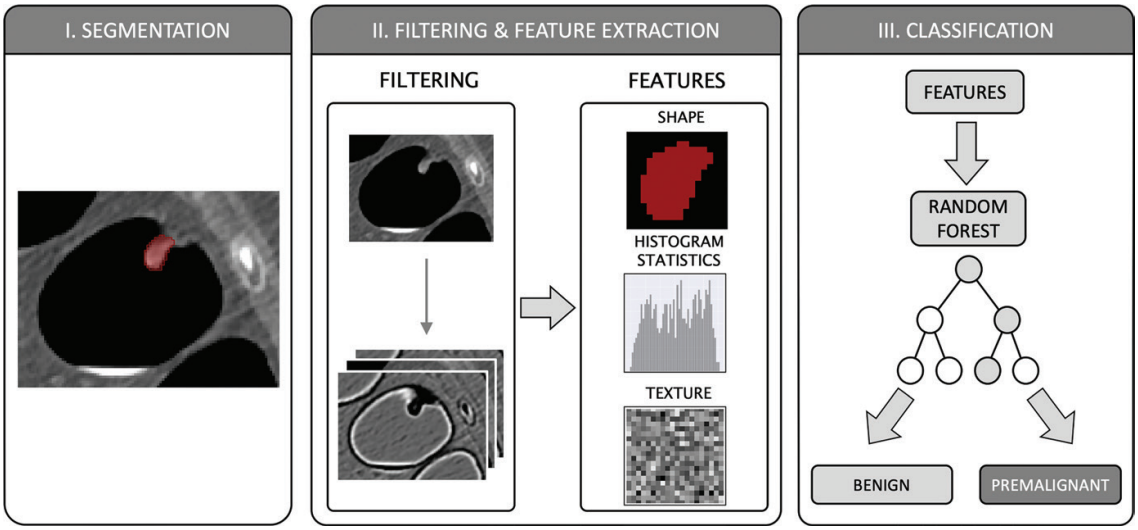

Figure 4: The radiomics workflow comprised three steps: manual segmentation of colorectal polyps in multiplanar two-dimensional CT colonography images; image filtering and feature extraction characterizing shape, histogram statistics, or texture; and feature-based training of a random forest classification algorithm to differentiate between benign and premalignant colorectal polyps according to the histopathologic reference standard.

of the 198 features used for training and testing the random forest. A detailed explanation of feature importance assessment is described in Appendix E1 (online).

During all stages of the analysis, the training and external test data sets were kept separate. The machine learning models were trained on the training set and evaluated on the test set. The statistical analysis was implemented in Python, and the entire code was made publicly available on the development platform Github *(https://github.com/pwesp/random-forest-polyp-classification)*.

Results

Training Set

Of 311 consecutively enrolled asymptomatic adults who underwent same-day CT colonography and OC, two (1%) had to be excluded because of withdrawal from the trial after CT colonography and two (1%) had to be excluded because of incomplete OC, as reported previously (9). In this retrospective analysis, 201 of 307 (65%) screening participants without histopathologically confirmed colorectal polyps were excluded. Of the resulting 106 colorectal cancer screening participants with histopathologically confirmed polyps, 43 (41%) were excluded because of incomplete or missing CT colonography imaging data. Of 164 colorectal polyps detected in OC, 57 (35%) were excluded because of retrospectively uncertain localization in CT colonography, retrospectively equivocal correlation to the histopathologic reference standard, or both. Thirty-five of the 57 (61%) excluded polyps were benign, 22 of 57 (39%) were premalignant, and further details are presented in Table E1 (online). Consensus reading due to divergent reading results was performed in five of 107 (5%) polyps. A total of 107 colorectal polyps were analyzed in 63 patients (mean age, 63 years ± 8; 40 men) comprising 169 segmentations of polyps

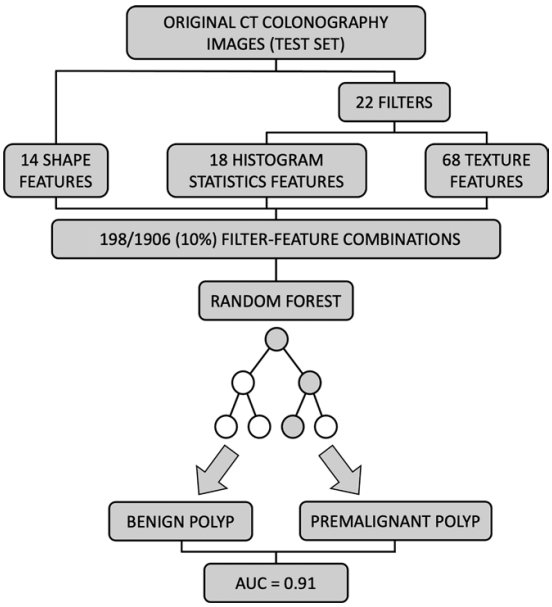

Figure 5: External validation of the trained random forest model. A total 198 of 1906 (10%) feature filter combinations were extracted from original CT colonography images of the test set using different image filters (n = 22) and image features characterizing shape (n = 14), histogram statistics (n = 18), or texture (n = 68). On the basis of these filter feature combinations, the trained random forest classifier was used to predict the colorectal polyp class label (benign vs premalignant). Prediction performance was quantified using area under the receiver operating characteristic curve (AUC).

(91 of 169 [54%] in the supine position, 78 of 169 [46%] in the prone position), as presented in Figure 1 and Table E2 (online). Of 169 polyp segmentations, 24 (14%) were 5 mm or smaller,

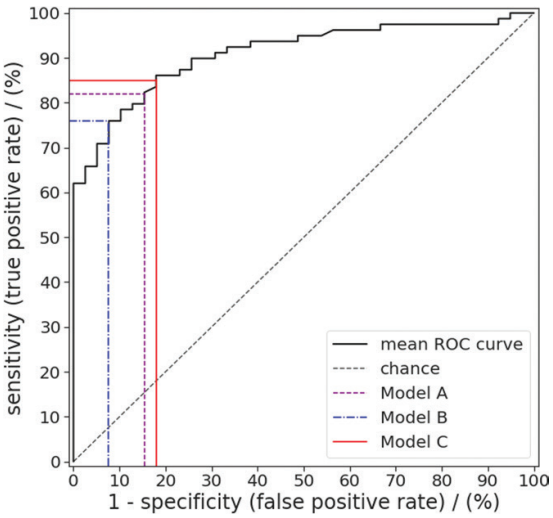

Figure 6: Receiver operating characteristic (ROC) curve for random forest predictions of colorectal polyp class (benign vs premalignant) for the polyps in the external test set. Sensitivity and specificity were evaluated for a default threshold value of 0.5 (model A), a threshold value that maximized the Youden index (model B), and a threshold resulting in the highest possible specificity while achieving a sensitivity of at least 85% (model C).

67 (40%) measured between 6 and 9 mm, and 78 of 169 (46%) were 10 mm or larger (maximum three-dimensional diameter measured through polyp segmentation). Eighty-three of 169 (49%) polyp segmentations were classified as benign (ie, hyperplastic polyp or regular mucosa), of which 16 of 83 (19%) were 5 mm or smaller, 49 of 83 (59%) were between 6 and 9 mm, and 18 of 83 (22%) were 10 mm or larger. Eighty-six of 169 (51%) polyp segmentations were classified as premalignant (ie, adenoma), of which eight of 86 (9%) were 5 mm or smaller, 18 of 86 (21%) were between 6 and 9 mm, and 60 of 86 (70%) were 10 mm or larger, as shown in Table 1. Polyp segmentation morphologies according to the size categories 5 mm or smaller, 6–9 mm, and 10 mm or larger are presented in Table 2.

External Test Set

A total of 53 of 194 (27%) colorectal polyps detected in OC were not clearly identifiable at CT colonography because of poor bowel distention or insufficient cleansing or tagging and were consequently excluded. Sixty-four of 194 (33%) polyps detected in OC were excluded because of retrospectively uncertain identification at CT colonography, equivocal correlation to the histopathologic reference standard, or both. A total of 58 of 117 (50%) excluded polyps were benign, whereas 59 of 117 (50%) were premalignant; further details are presented in Table E1 (online). Consensus reading due to divergent reading results was performed in five of 77 (7%) polyps. A total of 77 colorectal polyps was analyzed in 59 patients comprising 118 segmentations of polyps (56 of 118 [47%] in the supine position, 62 of 118 [53%] in the prone position). Nine of 118 (8%) polyp segmentations were 5 mm or smaller, 56 of 118 (47%) were between 6 and 9 mm, and 53 of 118 (45%) were 10 mm or larger, measuring the maximum three-dimensional diameter through polyp segmentation. Of 118, 39 (33%) polyp segmentations were classified as benign (ie, hyperplastic polyp or regular mucosa), of which eight (21%) were 5 mm or smaller, 26 (67%) were between 6 and 9 mm, and five (13%) were 10 mm or larger. Of the 118 polyp segmentations, 79 (67%) were classified as premalignant (ie, adenoma), of which one (1%) was 5 mm or smaller, 30 (38%) were between 6 and 9 mm, and 48 (61%) were 10 mm or larger, as shown in Table 1. Polyp segmentation morphologies according to the size categories 5 mm or smaller, 6–9 mm, and 10 mm or larger are presented in Table 2.

Statistical Analysis of the Test Set

Machine learning predictions of polyp class (benign vs premalignant) for all polyps in the external test set were derived from a random forest model fitted to the training set, yielding an overall AUC of 0.91 (95% CI: 0.85, 0.96). Sensitivity and specificity for a default threshold value of 0.5 (model A) were 82% (95% CI: 74, 91) (65 of 79) and 85% (95% CI: 72, 95) (33 of 39), respectively. Sensitivity and specificity for a threshold value of 0.53, which maximized the Youden index (model B), were 76% (95% CI: 64, 94) (60 of 79) and 92% (95% CI: 76, 100) (36 of 39), respectively. Sensitivity and specificity for a threshold value of 0.48, which resulted in the highest possible specificity while achieving a sensitivity of at least 85% (model C), were 85% (95% CI: 85, 89) (67 of 79) and 82% (95% CI: 61, 95) (32 of 39), respectively (Fig 6).

In subgroup analyses of the external test, which only contained polyps of a certain size category, machine learning predictions of polyp class (benign vs premalignant) yielded an AUC of 0.87 in the size category 6–9 mm and 0.90 in the size category 10 mm or larger. Because current guidelines recommend OC-guided resection for colorectal polyps 6 mm or larger (United States Multi-Society Task Force on Colorectal Cancer, European Society of Gastrointestinal Endoscopy, and European Society of Gastrointestinal and Abdominal Radiology) the number of polyps 5 mm or smaller with available histopathologic reports in the external test set (nine polyp segmentations) was not sufficient to provide a reliable AUC for this size category (13,14).

In a subgroup analysis of the external test set, in which only a single segmentation of each colorectal polyp was included, machine learning predictions of polyp class (benign vs premalignant) yielded an AUC of 0.90 for supine segmentations only and of 0.90 for prone segmentations only, respectively.

In an intrareader variability analysis, for which 29 of 118 (25%) polyp segmentations from the test set were segmented again by the same reader (Dice score, 0.80), random forest model predictions (benign vs premalignant) yielded an AUC of 0.87. In an interreader variability analysis, for which 29 of 118 (25%) polyp segmentations from the test set were segmented again by a second reader (Dice score, 0.77), random forest model predictions yielded an AUC of 0.91.

Table 1: Colorectal Polyp Segmentation Characteristics according to the Histopathologic Report of the Training Set and External Test Set

	No. of Polyp Segmentations		
Histopathologic Category	Training Set (n = 169)	External Test Set (n = 118)	Classification
Regular mucosa	3 (2)	9 (8)	Benign
Hyperplastic polyp	78 (46)	30 (25)	Benign
Lipomatous polyp	2 (1)	0 (0)	Benign
Tubular adenoma	57 (34)	49 (42)	Premalignant
Tubulovillous adenoma	16 (9)	26 (22)	Premalignant
Villous adenoma	8 (5)	0 (0)	Premalignant
Serrated adenoma	4 (2)	0 (0)	Premalignant
Adenocarcinoma	1 (1)	4 (3)	Premalignant

Note.—Data in parentheses are percentages. For study purposes only, a small number of adenocarcinoma segmentations were included in the premalignant group.

Table 2: Colorectal Polyp Segmentation Morphology in Three Size Categories of Both the Training Set and External Test Set

	No. of Polyp Segmentations					
	Training Set			External Test Set		
Morphologic Category	≤5 mm	6–9 mm	≥10 mm	≤5 mm	6–9 mm	≥10 mm
Pedunculated	0/169 (0)	5/169 (3)	32/169 (19)	0/118 (0)	1/118 (1)	17/118 (14)
Sessile	23/169 (14)	57/169 (34)	43/169 (25)	9/118 (8)	52/118 (44)	31/118 (26)
Flat	1/169 (1)	5/169 (3)	2/169 (1)	0/118 (0)	3/118 (3)	1/118 (1)
Carcinomatous	0/169 (0)	0/169 (0)	1/169 (1)	0/118 (0)	0/118 (0)	4/118 (3)

Note.—Data are proportions, and data in parentheses are percentages. Th e three size categories include the maximum three-dimensional diameter measured through polyp segmentation.

Statistical Analysis of the Training Set and Model Introspection

Polyp class predictions (benign vs premalignant) for out-of-bag training set samples yielded an AUC of 0.88. The 10 (of 198 [5%]) most important image features responsible for random forest classifications in this study were deduced from the trained random forest model and are shown in Table 3. Their relative feature importance added up to 23.8%. Seven features characterized texture (ranks 2, 3, 6–10), two features quantified first-order gray level histogram statistics (ranks 1 and 5), and one feature assessed size (rank 4). The remaining 188 (95%) features had a combined relative feature importance of 76.2%.

In detail, among the image features used for training the random forest model and making predictions, the size-measuring feature "original_shape_LeastAxisLength" was ranked as the fourth most important feature, with a relative feature importance of 2.6%, as shown in Table 3. It measured the smallest axis lengths of polyp segmentation–enclosing ellipsoids based on a principal component analysis.

Discussion

CT colonography does not enable a definite differentiation between benign and premalignant colorectal polyps, which would be crucial for individual risk stratification and therapy guidance. Consequently, in this proof-of-concept study, we investigated the noninvasive machine learning–based differentiation of benign and premalignant polyps in CT colonography data sets of an asymptomatic average-risk colorectal cancer screening cohort over 50 years of age. In correlation to the histopathologic reference standard, machine learning–based image analysis enabled robust differentiation of benign and premalignant CT colonography–detected colorectal polyps with an area under the receiver operating characteristic curve (AUC) of 0.91, even for small polyps in the size category of 6–9 mm with an AUC of 0.87. External validation of the technique in a large North American multicenter trial patient cohort demonstrated robustness and reproducibility of our method, despite data sets acquired by various scanner types and heterogeneous CT colonography imaging protocols (19–21). The feature importance analysis showed that a size-measuring image feature was in fourth place, but interestingly it was the only feature assessing size among the 10 most important image features for decision making in correlation to the histopathologic reference standard. The other nine image features characterized the distribution and texture of gray levels.

Our results are in line with pioneering studies by Aman et al and Song et al (30,31). Aman et al showed, in a data set of 97 polyps, that the differentiation of benign from premalignant polyps in machine learning–assisted CT colonography analysis using content-based image retrieval achieved a significantly higher (P = .048) AUC of 0.76 as opposed to the polyp size–only approach, with an AUC of 0.66 (30). Song et al reached an AUC of 0.85 for the machine learning–assisted differentiation of benign from premalignant colorectal polyps in a data set of 148 colorectal polyps, of which 35 were benign and 113 were premalignant (31). However, no external validation was performed to ensure the reliability in different scanner and protocol settings in these studies. In addition, Aman et al did not provide details on polyp size, and Song et al only included polyps 8 mm or larger (30,31). Our study adds to the field, as polyps smaller than 8 mm were analyzed, the machine learning algorithm was trained on a balanced training set, a standardized set of Pyradiomics image features was used, and validation on an external multicenter test set was performed.

Table 3: Ten Most Important Pyradiomics Image Features Responsible for Random Forest Model Classifications

Rank	Feature	Relative Importance (%)
1	log-sigma-2–0-mm-3D_firstorder_10Percentile	3.7
2	wavelet-LLH_glcm_Idmn	3.3
3	square_gldm_SmallDependenceLowGrayLevelEmphasis	3.2
4	original_shape_LeastAxisLength	2.6
5	original_firstorder_90Percentile	2.4
6	original_glcm_Imc2	2.2
7	wavelet-HLH_gldm_LargeDependenceHighGrayLevelEmphasis	1.9
8	wavelet-LLH_glrlm_LongRunLowGrayLevelEmphasis	1.5
9	exponential_glszm_SmallAreaLowGrayLevelEmphasis	1.5
10	gradient_glcm_Contrast	1.5

Note.—The relative feature importance added up to 23.8%. Seven features characterized texture (ranks 2, 3, 6–10), two features quantified first-order gray level histogram statistics (ranks 1 and 5), and one feature assessed size (rank 4). The remaining 188 (of 198 total) (95%) features had a combined relative feature importance of 76.2%. Filter abbreviations, feature category abbreviations, and feature abbreviations are presented in Tables E5–E7 (online).

Reproducibility between research groups and between image data sets acquired with different scanners using varying acquisition protocols pose a great challenge for machine learning–assisted image analysis (32,33). To address this issue and to ensure reproducibility of our machine learning model, only the standardized radiomics features available in the open-source Python package Pyradiomics were used (23), which is compliant with the Image Biomarker Standardization Initiative (34). The effect of intra- and interreader variability on our machine learning algorithm was investigated. To reduce the effect of varying CT colonography protocols, all CT colonography scans were resampled prior to feature extraction. Importantly, an external multicenter test set comprising CT colonography images acquired with multiple CT scanners from several vendors using varying CT colonography protocols was used to assess the performance of our radiomics-based machine learning differentiation of benign and premalignant colorectal polyps (19–21).

Brenner et al showed that OC cancer screening reduces the risk of colorectal cancer (7,35). However, these very successful OC screening programs had a low participation rate, at approximately 15%–20% (36–38). Colorectal cancer screening programs using noncathartic CT colonography and full cathartic preparation CT colonography show a higher participation rate compared with OC: 982 of 2920 (34%) for noncathartic CT colonography versus 1276 of 5924 (22%) for OC and 612 of 2430 (25%) for full cathartic preparation CT colonography versus 153 of 1036 (15%) for OC, respectively (37,38). Adding machine learning–assisted image analysis to conventional radiologic image reading could further improve the clinical importance of CT colonography–based colorectal cancer screening by allowing for a more precise selection of patients eligible for subsequent OC-guided polypectomy. Furthermore, machine learning–assisted CT colonography image analysis potentially provides the gastroenterologist with a road map of premalignant polyps eligible for OC-guided resection, particularly in patients with a high burden of colorectal polyps, such as in familial adenomatous polyposis.

Our study had limitations. The sample size was small, with 107 histopathologically confirmed colorectal polyps in the training set and 77 histopathologically confirmed colorectal polyps in the external test set. Every polyp securely identifiable in CT colonography data sets and unequivocally assignable to the corresponding histopathologic report was segmented. In the retrospective setting, however, a substantial number of polyps detected with OC did not meet these criteria and had to be excluded. Consequently, a selection bias could not be completely ruled out, and our results are only applicable to polyps clearly detectable with CT colonography. To ensure reproducibility of our machine learning model, only the standardized radiomics features available in the open-source Python package Pyradiomics were used (23). A dedicated feature measuring polyp height was not included, which might have been important for classification decisions, as hyperplastic polyps are known to be flatter than adenomatous polyps in CT colonography (39). Each patient, including all corresponding polyps and segmentations, was part of either the training set or the external test set. No patient who was presented to the random forest model during training was presented to the model again during testing. In a subgroup analysis of the external test set, in which only a single segmentation of each colorectal polyp was included, machine learning predictions of polyp class (benign vs premalignant) yielded an AUC of 0.90 for supine segmentations only and 0.90 for prone segmentations only. However, correlations within multiple polyps of one patient or within multiple segmentations of one polyp cannot be ruled out.

In this proof-of-concept study with validation in an external multicenter test set, machine learning–assisted CT colonography analysis enabled the differentiation of benign and premalignant colorectal polyps. The present study provides a potential basis for future prospective studies with a larger number of patients to further examine the diagnostic performance of machine learning algorithms for the noninvasive analysis of CT colonography–detected polyps.

Author contributions: Guarantors of integrity of entire study, S.G., P.W., M.I., P.M.K.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, S.G., P.W., C.S., T.K., C.C.C., J.R., M.I., P.M.K.; clinical studies, S.G., A.G., S.M., C.S., T.K., P.M.K.; statistical analysis, P.W., M.I.; and manuscript editing, S.G., P.W., A.G., C.S., C.C.C., J.R., M.I., P.M.K.

Disclosures of Conflicts of Interest: S.G. disclosed no relevant relationships. **P.W.** disclosed no relevant relationships. **A.G.** disclosed no relevant relationships. **S.M.** disclosed no relevant relationships. **C.S.** disclosed no relevant relationships. **T.K.** disclosed no relevant relationships. **C.C.C.** disclosed no relevant relationships. **J.R.** disclosed no relevant relationships. **M.I.** disclosed no relevant relationships. **P.M.K.** disclosed no relevant relationships.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70(1):7–30.
2. Ferlay J, Colombet M, Soerjomataram I, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer 2018;103:356–387.
3. Kumar V, Abbas AK, Aster JC, Robbins SL. Robbins basic pathology. Philadelphia, Pa: Elsevier/Saunders, 2013.
4. Mandel JS, Bond JH, Church TR, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. N Engl J Med 1993;328(19):1365–1371.
5. Winawer SJ, Zauber AG, Ho MN, et al. Prevention of colorectal cancer by colonoscopic polypectomy. N Engl J Med 1993;329(27):1977–1981.
6. Zauber AG, Winawer SJ, O'Brien MJ, et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med 2012;366(8):687–696.
7. Brenner H, Stock C, Hoffmeister M. Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. BMJ 2014;348:g2467.
8. Brenner H, Altenhofen L, Kretschmann J, et al. Trends in adenoma detection rates during the first 10 years of the German Screening Colonoscopy Program. Gastroenterology 2015;149(2):356–66.e1.
9. Graser A, Stieber P, Nagel D, et al. Comparison of CT colonography, colonoscopy, sigmoidoscopy and faecal occult blood tests for the detection of advanced adenoma in an average risk population. Gut 2009;58(2):241–248.
10. Kim DH, Pickhardt PJ, Taylor AJ, et al. CT colonography versus colonoscopy for the detection of advanced neoplasia. N Engl J Med 2007;357(14):1403–1412.
11. Atkin W, Dadswell E, Wooldrage K, et al. Computed tomographic colonography versus colonoscopy for investigation of patients with symptoms suggestive of colorectal cancer (SIGGAR): a multicentre randomised trial. Lancet 2013;381(9873):1194–1202.
12. Pooler BD, Kim DH, Weiss JM, Matkowskyj KA, Pickhardt PJ. Colorectal polyps missed with optical colonoscopy despite previous detection and localization with CT colonography. Radiology 2016;278(2):422–429.
13. Rex DK, Boland CR, Dominitz JA, et al. Colorectal cancer screening: recommendations for physicians and patients from the U.S. Multi-Society Task Force on Colorectal Cancer. Gastroenterology 2017;153(1):307–323.
14. Laghi A, Neri E, Regge D. Editorial on the European Society of Gastrointestinal Endoscopy (ESGE) and European Society of Gastrointestinal and Abdominal Radiology (ESGAR) guideline on clinical indications for CT colonography in the colorectal cancer diagnosis. Radiol Med (Torino) 2015;120(11):1021–1023.
15. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278(2):563–577.
16. Parekh V, Jacobs MA. Radiomics: a new application from established techniques. Expert Rev Precis Med Drug Dev 2016;1(2):207–226.
17. 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):1303–1322.
18. 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):2058–2067.
19. CT Colonography. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.NWTESAY1. Last modified June 3, 2020. Accessed November 17, 2019.
20. Johnson CD, Chen MH, Toledano AY, et al. Accuracy of CT colonography for detection of large adenomas and cancers. N Engl J Med 2008;359(12):1207–1217.
21. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26(6):1045–1057.
22. Nolden M, Zelzer S, Seitel A, et al. The Medical Imaging Interaction Toolkit: challenges and advances: 10 years of open-source development. Int J CARS 2013;8(4):607–620.
23. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–e107.
24. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html.
25. Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
26. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, NY: Springer, 2009.
27. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–35.
28. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19(9):1141–1164.
29. Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–140.
30. Aman JM, Yao J, Summers RM. Prediction of polyp histology on CT colonography using content-based image retrieval. In: Karssemeijer N, Summer RM, eds. Proceedings of SPIE: medical imaging 2010—computer-aided diagnosis. Vol 7624. Bellingham, Wash: International Society for Optics and Photonics, 2010; 76240D.
31. Song B, Zhang G, Lu H, et al. Volumetric texture features from higher-order images for diagnosis of colon lesions via CT colonography. Int J CARS 2014;9(6):1021–1031.
32. 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):221–228.
33. Meyer M, Ronald J, Vernuccio F, et al. Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings. Radiology 2019;293(3):583–591.
34. 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):328–338.
35. Brenner H, Chang-Claude J, Jansen L, Knebel P, Stock C, Hoffmeister M. Reduced risk of colorectal cancer up to 10 years after screening, surveillance, or diagnostic colonoscopy. Gastroenterology 2014;146(3):709–717.
36. van der Meulen MP, Lansdorp-Vogelaar I, Goede SL, et al. Colorectal cancer: cost-effectiveness of colonoscopy versus CT colonography screening with participation rates and costs. Radiology 2018;287(3):901–911.
37. Stoop EM, de Haan MC, de Wijkerslooth TR, et al. Participation and yield of colonoscopy versus non-cathartic CT colonography in population-based screening for colorectal cancer: a randomised controlled trial. Lancet Oncol 2012;13(1):55–64.
38. Sali L, Mascalchi M, Falchini M, et al. Reduced and full-preparation CT colonography, fecal immunochemical test, and colonoscopy for population screening of colorectal cancer: a randomized trial. J Natl Cancer Inst 2015;108(2):djv319.
39. Summers RM, Liu J, Yao J, Brown L, Choi JR, Pickhardt PJ. Automated measurement of colorectal polyp height at CT colonography: hyperplastic polyps are flatter than adenomatous polyps. AJR Am J Roentgenol 2009;193(5):1305–1310.

5 | Publication II

European Radiology (2022) 32:4749–4759
https://doi.org/10.1007/s00330-021-08532-2

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Deep learning in CT colonography: differentiating premalignant from benign colorectal polyps

Philipp Wesp[1] · Sergio Grosu[1] · Anno Graser[2] · Stefan Maurus[1] · Christian Schulz[3] · Thomas Knösel[4] · Matthias P. Fabritius[1] · Balthasar Schachtner[1,5] · Benjamin M. Yeh[6] · Clemens C. Cyran[1] · Jens Ricke[1] · Philipp M. Kazmierczak[1] · Michael Ingrisch[1]

Received: 18 August 2021 / Revised: 6 December 2021 / Accepted: 20 December 2021 / Published online: 26 January 2022

Abstract

Objectives To investigate the differentiation of premalignant from benign colorectal polyps detected by CT colonography using deep learning.

Methods In this retrospective analysis of an average risk colorectal cancer screening sample, polyps of all size categories and morphologies were manually segmented on supine and prone CT colonography images and classified as premalignant (adenoma) or benign (hyperplastic polyp or regular mucosa) according to histopathology. Two deep learning models SEG and noSEG were trained on 3D CT colonography image subvolumes to predict polyp class, and model SEG was additionally trained with polyp segmentation masks. Diagnostic performance was validated in an independent external multicentre test sample. Predictions were analysed with the visualisation technique Grad-CAM++.

Results The training set consisted of 107 colorectal polyps in 63 patients (mean age: 63 ± 8 years, 40 men) comprising 169 polyp segmentations. The external test set included 77 polyps in 59 patients comprising 118 polyp segmentations. Model SEG achieved a ROC-AUC of 0.83 and 80% sensitivity at 69% specificity for differentiating premalignant from benign polyps. Model noSEG yielded a ROC-AUC of 0.75, 80% sensitivity at 44% specificity, and an average Grad-CAM++ heatmap score of ≥ 0.25 in 90% of polyp tissue.

Conclusions In this proof-of-concept study, deep learning enabled the differentiation of premalignant from benign colorectal polyps detected with CT colonography and the visualisation of image regions important for predictions. The approach did not require polyp segmentation and thus has the potential to facilitate the identification of high-risk polyps as an automated second reader.

Key Points

- *Non-invasive deep learning image analysis may differentiate premalignant from benign colorectal polyps found in CT colonography scans.*
- *Deep learning autonomously learned to focus on polyp tissue for predictions without the need for prior polyp segmentation by experts.*
- *Deep learning potentially improves the diagnostic accuracy of CT colonography in colorectal cancer screening by allowing for a more precise selection of patients who would benefit from endoscopic polypectomy, especially for patients with polyps of 6–9 mm size.*

Keywords Colonography · Computed tomographic · Colonic polyp · Deep learning · Early detection of cancer

Philipp Wesp and Sergio Grosu contributed equally to this work

✉ Philipp Wesp
philipp.wesp@med.uni-muenchen.de

[1] Department of Radiology, University Hospital, LMU Munich, Marchioninistraße 15, 81377 Munich, Germany

[2] Radiologie München, Burgstraße 7, 80331 Munich, Germany

[3] Department of Medicine II, University Hospital, LMU Munich, Marchioninistraße 15, 81377 Munich, Germany

[4] Department of Pathology, University Hospital, LMU Munich, Marchioninistraße 15, 81377 Munich, Germany

[5] Comprehensive Pneumology Center (CPC-M), Member of the German Center for Lung Research (DZL), Max-Lebsche-Platz 31, 81377 Munich, Germany

[6] Department of Radiology and Biomedical Imaging, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94117, USA

Abbreviations

AUC	Area under the curve
CNN	Convolutional neural network
MITK	Medical Imaging Interaction Toolkit
OC	Optical colonoscopy
PEG	Polyethylene glycol solution
ROC	Receiver operating characteristics
TCIA	The Cancer Imaging Archive

Introduction

Colorectal cancer is one of the three most frequent cancer-related causes of death among men and women [1]. However, its mortality and incidence can be significantly decreased by early detection of precancerous adenomatous polyps which grow over several years [2–5]. Screening methods such as immunochemical faecal occult blood test and optical colonoscopy (OC) are proven to reduce mortality from colorectal cancer, particularly since clinical symptoms are often non-specific or absent [6, 7].

A non-invasive screening method for colorectal cancer is computed tomography (CT) colonography. For the detection of colorectal polyps ≥ 6 mm, the sensitivity of CT colonography is comparable to OC [8–10]. Computer-aided detection (CAD) algorithms can reduce the number of missed colorectal polyps at CT colonography when used as a second reader [11, 12].

However, conventional CT colonography does not allow a clear distinction between benign and premalignant colorectal polyps, which would be essential for individual risk stratification and therapy management. Premalignant adenomatous polyps require endoscopic resection, whereas benign findings of hyperplastic polyps avoid unnecessary interventions. As polyp size is the only surrogate indicator of the likelihood of malignancy at CT colonography, current guidelines recommend the resection of colorectal polyps ≥ 6 mm detected in CT colonography (European Society of Gastrointestinal and Abdominal Radiology, United States Multi-Society Task Force on Colorectal Cancer) [13, 14].

First studies have shown that machine learning–based CT colonography using radiomics may allow non-invasive differentiation of benign and premalignant colorectal polyps [15, 16]. These radiomics approaches consist of three steps. First, segmentation of the region-of-interest in the medical image, i.e. the polyp in the CT colonography scan. Second, extraction of radiomics features for the segmented regions. Third, machine learning analysis of the extracted features to predict polyp character. Especially the first step of polyp segmentation, which has been performed manually by experts, is a large barrier for the potential integration of these approaches into the clinical routine and prevents fully automated polyp classification. In addition, the interpretability of these approaches is limited to the importance of individual radiomics features. Deep learning could potentially overcome these challenges and thereby substantially reduce the gap to clinical applicability for machine learning–based polyp classification in CT colonography.

Deep learning–based image classification using convolutional neural networks (CNNs) does not require prior segmentation of the region-of-interest and has proven to be an efficient method in automated image analysis, providing a powerful tool for tumour detection and classification in oncologic imaging [17]. In the first step of a deep learning approach, a localisation of the polyp is sufficient. In the second step, a deep learning model can directly predict polyp character using a small subvolume of the CT colonography image around the localisation. Additionally, CNNs can be exploited to visualise regions in the input image that are potentially important for model predictions to achieve improved model interpretability [18].

Therefore, the aim of this study was to establish the differentiation of premalignant (i.e. adenoma) and benign (i.e. hyperplastic polyp or regular mucosa) colorectal polyps in CT colonography using deep learning.

Materials and methods

Training set

This study was approved by the institutional review board and the requirement for written informed consent was waived. It is a retrospective analysis of CT colonography images from a previously published prospective colorectal cancer screening cohort of an asymptomatic, average risk population over 50 years of age who underwent same-day OC and CT colonography [8]. Exclusion criteria of the previously published colorectal cancer screening cohort were signs of colonic illnesses such as abdominal pain, relevant changes in stool frequency, diarrhoea, melaenic stools, and haematochezia as well as positive family history for colorectal cancer, hereditary colorectal cancer syndromes, inflammatory bowel disease, severe cardiovascular or pulmonary disease, body weight > 150 kg, and prior OC within the last 5 years. Only participants with histopathologically confirmed findings corresponding to CT colonography findings were included in the present study (Fig. 1).

CT colonography in the training set

CT colonography bowel preparation is described in the Supplemental Material. CT colonography images were acquired on a 64-channel multidetector row scanner (Siemens Somatom Sensation 64, Siemens Healthineers) at 0.6 mm collimation and reconstructed using a standard soft

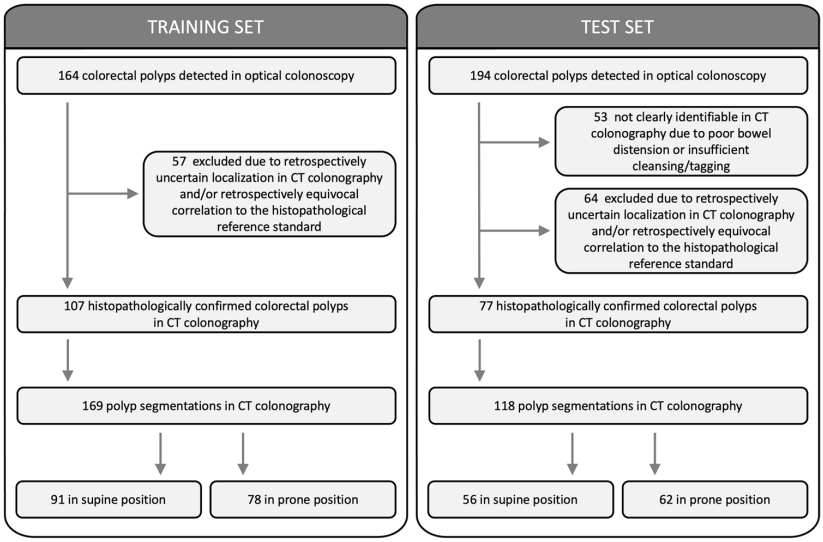

Fig. 1 Flow diagram of the training set and the external test set

tissue kernel at a slice thickness of 0.75 mm and 0.5 mm reconstruction increment. Tube voltage was 120 kVp at tube current–time product reference values of 70 mAs in supine and 30 mAs in prone position using automatic tube current adaption. Mean radiation dose for CT colonography was 4.5 (0.6) mSv. For bowel distension, room air or CO_2 was insufflated through a rectal tube. No intravenous contrast agent was administered. The CT colonography protocol was described in detail before [8].

External test set

CT colonography datasets from a North American multicentre CT colonography screening trial, publicly available via The Cancer Imaging Archive (TCIA), served as an external test set [19–21]. The external test set comprised multicentre CT colonography images acquired on various CT scanners from different vendors (Siemens Healthineers; Philips Healthcare; GE Healthcare Systems; Canon Medical Systems) with varying scanning protocols. Polyps were only included if histopathologic reports were available.

Polyp segmentation

Prospective polyp detection and polyp matching are described in the Supplemental Material. All readers were informed about polyp size and colon segment in which polypectomy was performed. Histopathological polyp class was blinded for all readers. Colorectal polyps were manually segmented in multiplanar 2D CT colonography images by a board-certified radiologist (8 years of experience in CT colonography imaging; completed a specialised hands-on workshop on CT colonography) and two radiology residents (3 years of experience in CT colonography imaging; one completed a specialised hands-on workshop on CT colonography) in equal amounts, as described in detail beforehand [16]. For exact retrospective polyp re-detection, 2D and virtual fly-through 3D CT colonography reconstructions were used (Fig. 2). Colorectal polyps that could not be clearly identified in CT colonography and/or unequivocally assigned to the corresponding histopathological report were excluded. A consensus reading was performed in case of divergent reading results. Consensus was reached when all three readers agreed on polyp localisation and segmentation. Each colorectal polyp was segmented in supine and prone position images, if confidently detectable in both positions. The CT colonography workflow of the dedicated post-processing software syngo.via versionVA30B (Siemens Healthineers) was used for polyp detection. The Medical Imaging Interaction Toolkit (MITK) Version 2018.04 (German Cancer Research Center — Division of Medical Image Computing) was used for polyp segmentation [22].

Histopathological reference standard

A colorectal polyp was considered benign if the corresponding histopathological report classified it as “regular mucosa” or “hyperplastic polyp”, premalignant if the corresponding histopathological report classified it as “tubular adenoma”, “tubulovillous adenoma”, or “villous adenoma”.

Solely for study purposes, 2 lesions with the histopathological classification “serrated adenoma” and 3 lesions with the histopathological classification “adenocarcinoma” (39 mm, 44 mm, and 75 mm) were included in the group

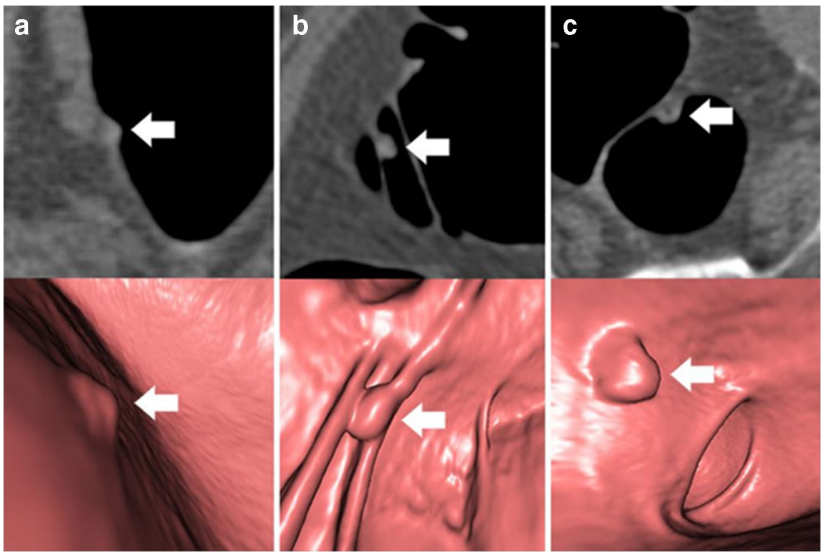

Fig. 2 **a-c** Colorectal polyps of the training set (indicated by arrows) in axial 2D CT colonography images (top row) and in the corresponding virtual fly-through 3D reconstructions (bottom row). **a** 7-mm hyperplastic polyp in the rectum of a 58-year-old woman. **b** 8-mm tubular adenoma in the transverse colon of a 74-year-old woman. **c** 9-mm tubulovillous adenoma in the rectum of a 67-year-old man

Table 1 Colorectal polyp segmentations in the training set and external test set class-divided according to the histopathological report

Histopathologic category	Number of polyp segmentations		Classification
	Training set	External test set	
Regular mucosa	3/169 (2%)	9/118 (8%)	Benign
Hyperplastic polyp	78/169 (46%)	30/118 (25%)	
Lipomatous polyp	2/169 (1%)	0/118 (0%)	
Tubular adenoma	57/169 (34%)	49/118 (42%)	Premalignant
Tubulovillous adenoma	16/169 (9%)	26/118 (22%)	
Villous adenoma	8/169 (5%)	0/118 (0%)	
Serrated adenoma	4/169 (2%)	0/118 (0%)	
Adenocarcinoma	1/169 (1%)	4/118 (3%)	

The adenocarcinoma segmentations were included in the premalignant group for study purposes only

premalignant. One polyp with the histopathological classification "lipomatous" was included in the group benign (Table 1).

Deep learning–based ensemble models

This study investigated two deep learning–based models, SEG and noSEG. Both models were ensembles, each consisting of 50 three-dimensional convolutional neural networks [23]. In each ensemble, the mean output of the 50 respective CNNs was used as model output. Ensembling was implemented to address the variance observed while training single CNNs. This variance was believed to be an effect of training set size — deep learning is typically applied on large datasets — and could not be eliminated with data augmentation. The CNNs used in both ensemble models were, apart from the input layer, identical (Fig. 3). CNNs in SEG expected inputs of size $50 \times 50 \times 50 \times 2$ (image + segmentation), CNNs in noSEG expected inputs of size $50 \times 50 \times 50 \times 1$ (image). A CNN from model noSEG is shown schematically in Fig. 4 and a detailed layer-by-layer description for CNNs from both models is provided in Table 2. Both models were implemented with Keras (version 2.4.3) [24], an open-source Python interface for neural networks. The open-source machine learning library TensorFlow (Google Brain, version 2.4.1) [25] was used as backend.

CNN training

Every CNN in each of the models was trained individually to predict the histopathological polyp class label (benign vs. premalignant). CNNs in SEG were trained with images and segmentations; CNNs in noSEG were trained with images exclusively (Fig. 3). Every CNN was trained with a different 80–20 train-validation split. In these splits, 80% of the data were randomly selected as training data to train the network, and the other 20% were used as validation data to monitor the training process. Training parameters included a stochastic gradient descent (SGD) optimiser, a learning rate of 0.01, a binary cross-entropy loss function, and a batch size of 8. Data augmentation, including random cropping, was used in the training data. The validation data was not augmented, but cropped to size $50 \times 50 \times 50$ around the polyp centre to match the input size. Early stopping was applied to automatically end the training process: If the AUC in the

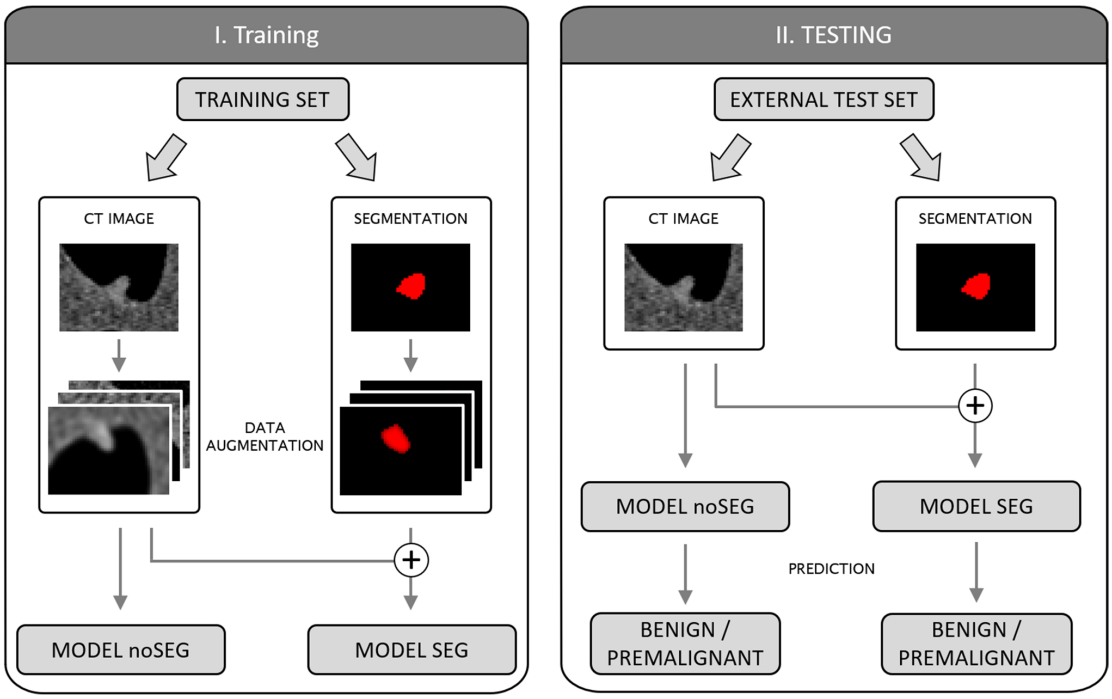

Fig. 3 Schematic illustration of model training (left) and testing (right). Training: Model noSEG was trained on augmented CT images of the training set, model SEG was trained on augmented CT images and manual polyp segmentation masks. Testing: Model noSEG predicted polyp class (benign vs. premalignant) on CT images of the independent external test set, and model SEG made predictions based on CT images and manual polyp segmentation masks

20%-validation set did not increase for 64 epochs, training was stopped and the weights from the epoch with the highest validation AUC were restored.

Statistical analysis of the external test set

The classification performance of the trained models SEG and noSEG was evaluated on the independent, external test set (Fig. 3). Model output scores were calculated as the arithmetic mean of the 50 individual output scores of the CNNs in each ensemble for each input image. The model output score was turned into a prediction using a classification threshold. The threshold was selected to yield a sensitivity of 80%. Classification performance was quantified using AUC, sensitivity, and specificity. For polyp size–based subgroup analyses, the maximum polyp diameter in three dimensions was calculated based on the polyp segmentation masks.

Visual explanation of model predictions

The gradient-based CNN visualisation technique GradCAM++ [18] provided visual explanations of predictions made for the test set by model noSEG (predictions based on input images exclusively). For each voxel in an input image, GradCAM++ calculated a class activation, ranging from 0.0 to 1.0, to visualise the correspondence with the model output score. GradCAM++ images for three selected polyps are shown in Fig. 5. In addition, we quantified how much attention the model noSEG paid to voxels labelled as "polyp", according to the manual polyp segmentation masks, during decision-making and calculated the percentage of voxels inside the manual polyp segmentation mask which had a GradCAM++ class activation of 0.25 or higher (Fig. 5).

The code for the statistical analysis was made publicly available on the development platform GitHub at https://github.com/pwesp/deep-learning-in-ct-colonography.

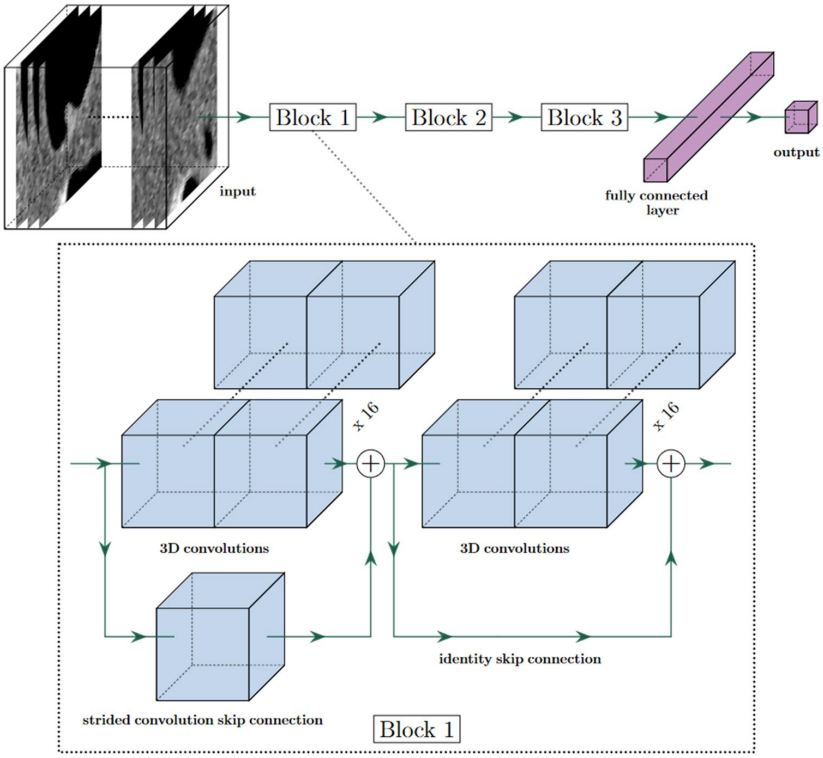

Fig. 4 Schematic illustration of the CNN architecture used in the ensemble models SEG and noSEG. First, the input (CT image for model noSEG, CT image and manual polyp segmentation mask for model SEG) propagates through three convolution blocks (blocks 1, 2, and 3), each consisting of two consecutive three-dimensional convolutions with an increasing number of filter kernels (block 1: 16 kernels, block 2: 32 kernels, block 3: 64 kernels) and skip connections. Afterwards, a fully connected layer mapped the information to the output neuron which holds the output score (0.0 = benign, 1.0 = pre-malignant)

Table 2 Layer-by-layer description of the CNNs used in the two ensemble models SEG and noSEG

Name	Layer	Filter kernel (shape, count)		Output size	
		Main branch	Shortcut	noSEG	SEG
in	Input	-		50×50×50×1	50×50×50×2
res1a	3D convolution	3×3×3, 16	3×3×3, 1	25×25×25×16	
res1b	3D convolution	3×3×3, 16	id	25×25×25×16	
add1	Add	-		25×25×25×16	
res2a	3D convolution	3×3×3, 32	3×3×3, 1	13×13×13×32	
res2b	3D convolution	3×3×3, 32	id	13×13×13×32	
add2	Add	-		13×13×13×32	
res3a	3D convolution	3×3×3, 64	3×3×3, 1	7×7×7×64	
res3b	3D convolution	3×3×3, 64	id	7×7×7×64	
add3	Add	-		7×7×7×64	
pool	Global average pooling	-		64	
drop	Dropout	-		64	
out	Fully connected layer	-		1	

The convolutional part of each network (up to layer "add3") consisted of a main branch, containing three-dimensional convolutions, and a shortcut branch, containing either a single convolution kernel for downscaling or an identity mapping ("id"). At each add layer ("add1", "add2", "add3"), the main branch and the shortcut branch were added. After add1 and add2, the images were split up again into main and shortcut branches

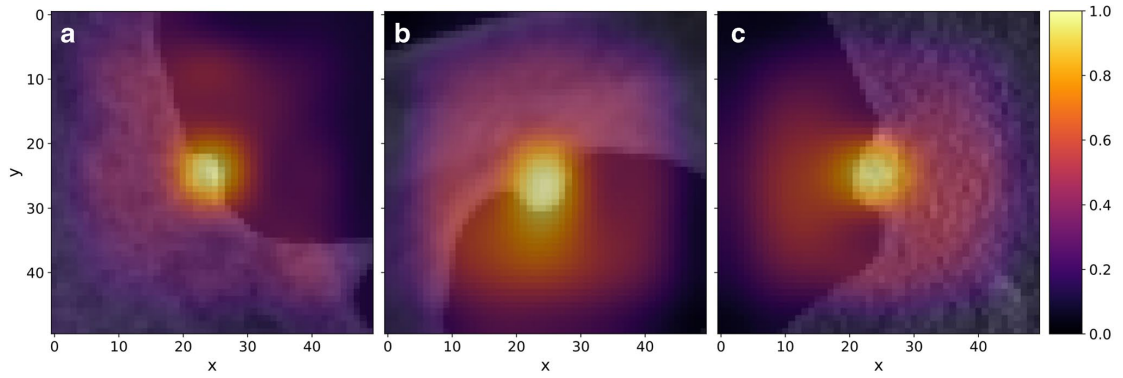

Fig. 5 GradCAM++ images of model noSEG for the inputs (**a**) 7-mm hyperplastic polyp, (**b**) 7-mm tubular adenoma, and (**c**) 9-mm tubulovillous adenoma from the test set superimposed with the respective 2D CT colonography images. Grad-CAM++ is a gradient-based explanation method for CNNs and was used to visualise the correspondence (0.0 = no correspondence, 1.0 = highest correspondence) of each image voxel with the prediction of the model noSEG (benign vs. premalignant polyp) [18]

Results

Training set

Of 311 consecutively enrolled adults undergoing same-day CT colonography and OC, 2 had to be excluded due to withdrawal from the trial after CT colonography and 2 because of incomplete OC, as reported previously [8]. Of 307 colorectal cancer screening participants of an average risk asymptomatic screening population, 201 participants without findings of histopathologically confirmed polyps were excluded. Of 106 participants with histopathologically confirmed polyps, 43 were excluded due missing or incomplete CT colonography datasets. Of 164 colorectal polyps detected in OC, 57 were excluded due to retrospectively equivocal assignment to the histopathological reference standard and/or retrospectively uncertain localisation in CT colonography, as described in detail previously [16]. Thirty-five of 57 excluded polyps were benign, and 22 of 57 were premalignant. Consensus reading was performed in 5 of 107 polyps. In total, 107 colorectal polyps with histopathological reference were evaluated in 63 patients (23 female; mean age: 63 ± 8 years) comprising 169 polyp segmentations in CT colonography images (91 in supine position and 78 in prone position). Eighty-six polyp segmentations were categorised as premalignant (adenoma), of which 8 were ≤ 5 mm, 18 between 6 and 9 mm, and 60 ≥ 10 mm, measuring the maximum 3D diameter of polyp segmentations. Eighty-three polyp segmentations were categorised as benign (hyperplastic polyp or regular mucosa), of which 16 were ≤ 5 mm, 49 between 6 and 9 mm, and 18 ≥ 10 mm.

External test set

Due to insufficient cleansing/tagging or poor bowel distension, 53 of 194 colorectal polyps detected in OC were not clearly identifiable in CT colonography. Sixty-four polyps were excluded due to retrospectively equivocal assignment to the histopathological reference standard and/or retrospectively uncertain localisation in CT colonography, as described in detail before [16]. Fifty-eight of 117 excluded polyps were benign, and 59 of 117 were premalignant. Consensus reading was performed in 5 of 77 polyps. In total, 77 colorectal polyps were analysed in 59 patients comprising 118 polyp segmentations (56 in supine position and 62 in prone position). Seventy-nine polyp segmentations were categorised as premalignant (adenoma), of which 1 was ≤ 5 mm, 30 between 6 and 9 mm, and 48 ≥ 10 mm. Thirty-nine polyp segmentations were categorised as benign (hyperplastic polyp or regular mucosa), of which 8 were ≤ 5 mm, 26 between 6 and 9 mm, and 5 ≥ 10 mm.

Statistical analysis of the external test set

On the independent, external test set, output scores from model SEG yielded an AUC of 0.83, and output scores from model noSEG yielded an AUC of 0.75. Model predictions for polyp class from model SEG yielded a sensitivity and specificity of 80% (63 of 79) and 69% (27 of 39) for a classification threshold of 0.27. noSEG predictions for polyp class yielded a sensitivity and specificity of 80% (63 of 79) and 44% (17 of 39) for a classification threshold of 0.36 (Fig. 6).

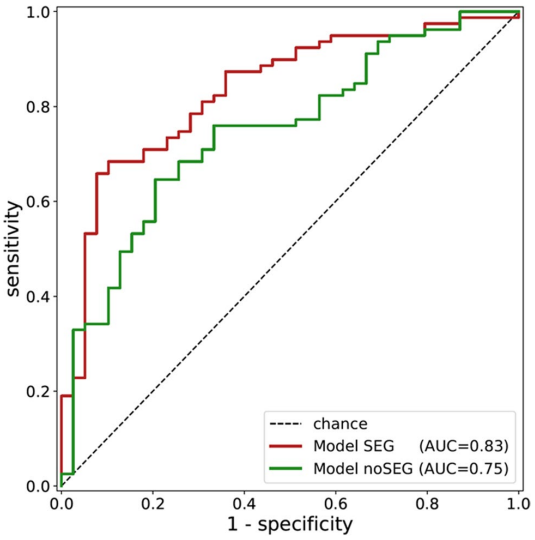

Fig. 6 Receiver operating characteristic (ROC) curve for deep learning predictions of polyp class (benign vs. premalignant) in the external test set from model SEG and model noSEG

Table 3 Class prediction accuracy of the two models SEG and noSEG on the external test set polyp segmentations for each histopathologic category

Histopathologic category	Model accuracy		Ground truth classification
	SEG	noSEG	
Regular mucosa	3/9 (33%)	7/9 (78%)	Benign
Hyperplastic polyp	21/30 (70%)	15/30 (50%)	
Lipomatous polyp	0/0	0/0	
Tubular adenoma	36/49 (73%)	37/49 (76%)	Premalignant
Tubulovillous adenoma	23/26 (88%)	23/26 (88%)	
Villous adenoma	0/0	0/0	
Serrated adenoma	0/0	0/0	
Adenocarcinoma	4/4 (100%)	3/4 (75%)	

The four adenocarcinoma segmentations were included in the premalignant group for study purposes only

Visual explanations of deep learning predictions were provided using the gradient-based CNN visualisation technique GradCAM++. The fraction of manual polyp segmentation mask voxels which had a GradCAM++ class activation of 0.25 or higher from model noSEG was 90% on average.

In size-based subgroup analyses of the external test set, model SEG yielded an AUC of 0.74 for polyps with a size between 6 and 9 mm and 0.72 for polyps ≥ 10 mm. Model noSEG yielded an AUC of 0.72 for polyps with a size between 6 and 9 mm and 0.74 for polyps ≥ 10 mm. As current guidelines recommend endoscopic resection for colorectal polyps ≥ 6 mm, the number of polyps ≤ 5 mm with available histopathologic classification in the external test set (9 polyp segmentations) was not sufficient to provide reliable results for this size category [13, 14].

In a further subgroup analysis of the external test set based on the histopathologic report (see Table 3), tubulovillous adenoma had the highest percentage of correctly classified cases (SEG: 23/26 (88%); noSEG: 23/26 (88%)), followed by tubular adenoma (SEG: 36/49 (73%); noSEG: 37/49 (76%)) and hyperplastic polyp (SEG: 21/30 (70%); noSEG: 15/30 (50%)).

Discussion

In this proof-of-concept study, we investigated the deep learning–based differentiation of premalignant and benign colorectal polyps in CT colonography datasets of an average-risk, asymptomatic colorectal cancer screening cohort of over 50 years of age. Deep learning–based image analysis allowed for the differentiation of benign and premalignant colorectal polyps with CT colonography with an AUC of 0.83. Even when manual polyp segmentations were not used for decision-making, deep learning reached an AUC of 0.75. External validation demonstrated robustness of the deep learning models, despite images acquired with heterogeneous CT colonography imaging protocols on various CT scanners [19–21]. Tubulovillous adenomas were classified with higher accuracy (88% each model) compared to less premalignant tubular adenoma (73% and 76%). This might indicate that, for premalignant polyps, the differentiation performance is increased with higher malignant potential of polyps.

The use of deep learning for the classification of colorectal polyps in CT colonography is not yet well established. In a pioneering study, Tan et al. investigated a deep learning–based classification of colorectal lesions > 30 mm detected in CT colonography in correlation to the histopathological reference standard [26]. Tubular adenoma, tubulovillous adenoma, and villous adenoma were labelled as benign ($N = 31$); adenocarcinomas were labelled as malignant ($N = 32$). In two-fold cross validation, a deep learning model trained on CT colonography images reached an AUC of 0.84 [26].

Our study adds to the literature, as we showed the ability of deep learning–based image classification at CT colonography to differentiate between adenomas (premalignant) and hyperplastic polyps (benign), considering that most colorectal cancers develop from adenomas and the incidence of colorectal cancer can be significantly decreased by early detection with subsequent resection

[2–4]. As we included polyps ≤ 9 mm ($N = 91$ images in the training set, $N = 65$ images in the external validation), our results show that small colorectal polyps can be classified as benign or premalignant using deep learning. Furthermore, we evaluated the performance of our deep learning–based models in an independent, external, multicentre test set.

Besides deep learning, classical machine learning methods have been used for colorectal polyp classification in CT colonography as part of a radiomics approach [15, 16]. Radiomics approaches typically consist of three steps: region-of-interest segmentation, radiomics feature extraction, machine learning prediction. In a previous analysis of this training and external test dataset using such a radiomics approach, a random forest machine learning model enabled the robust differentiation of benign and premalignant CT-colonography-detected colorectal polyps with an AUC of 0.91 [16]. The higher performance compared to deep learning (AUC of 0.84 and 0.75) can be attributed to the relatively small size of the training dataset. Deep learning typically requires larger amounts of data for successful training than classical machine learning methods like random forests [17, 27].

The present study provides additional value as, contrary to a radiomics approach, deep learning–based CT colonography image analysis did not require polyp segmentation. Merely a localisation of the polyp had to be provided. Additionally, deep learning models extract image features and make predictions at the same time, which leads to an approach with just two steps: localisation and deep learning prediction. This promises application in clinical routine, since polyp localisation would be more feasible compared to a thorough segmentation. Furthermore, it provides the basis for a fully automated CT colonography evaluation as the deep learning–based polyp classification could be combined with already established CAD algorithms for polyp detection [11, 12]. Additionally, the CNNs which made up the deep learning models enabled the visual interpretation of predictions. We used the gradient-based CNN visualisation technique GradCAM++ [18] to highlight regions in the input CT colonography image that were potentially relevant for decision-making. High activation in image regions that were manually labelled by radiologists to create polyp segmentation masks confirmed that model noSEG was capable of recognising autonomously which image voxels were important for decision-making, without the need for preidentification via polyp segmentation. In contrast, radiomics approaches typically allow to rank features according to their importance during training a classical machine learning model. However, the majority of radiomics features are second-order texture features which are difficult to interpret in a medical context.

Used as a second reader, deep learning–based CT colonography analysis could further increase the clinical impact of CT colonography–based colorectal cancer screening by enabling a more precise selection of patients who would profit from subsequent endoscopic polypectomy. Particularly considering that colorectal cancer screening programs using CT colonography showed higher participation rates compared to OC [28, 29]. Current guidelines recommend the resection of colorectal polyps ≥ 6 mm detected in CT colonography [13, 14]. One reason for this recommendation is that colonoscopic referral for polyps with a size of ≤ 5 mm at screening CT colonography has been shown to have very poor cost-effectiveness with $464,407 per life-year gained [30]. Furthermore, Pickhardt et al. demonstrated that the incremental cost-effectiveness ratio of colonoscopic referral for polyps with a size between 6 and 9 mm at CT colonography was $59,015 per life-year gained, compared to − $151 (cost savings per person) for polyps with a size of ≥ 10 mm [30]. By allowing the differentiation of premalignant from benign colorectal polyps, especially in the size category between 6 and 9 mm, deep learning–based CT colonography analysis could potentially increase the cost-effectiveness ratio of colonoscopic referral after CT colonography.

This study has limitations. The sample size was small. Every polyp securely identifiable in CT colonography and unequivocally assignable to the corresponding histopathological report was segmented. A substantial number of polyps detected in OC, however, had to be excluded. Therefore, the results of this study are only applicable to polyps clearly detectable in CT colonography and a selection bias cannot be fully ruled out. No polyp that was presented to a deep learning model during training was presented to the model again during testing. However, correlations within multiple segmentations of one polyp or within multiple polyps of one patient in model SEG cannot be ruled out. The prevalence of serrated adenomas in this study (1.6%) (2 out of 122 patients) was in agreement with the prevalence of serrated adenomas in a large-scale CT colonography screening study (1.4%) [31]. However, the number of serrated adenomas ($N = 2$) was not sufficient to provide reliable results for deep learning–based analysis of this category.

Conclusions

In this proof-of-concept study, deep learning–based analysis of CT colonography allowed differentiating premalignant from benign colorectal polyps in an external validation cohort corresponding to histopathology. Differentiation was possible, even when the model was provided only CT images and did not utilise expert polyp segmentation masks. Deep learning allowed for visual interpretability of the results

so that image regions potentially important for predictions could be analysed. Although the findings need to be validated in prospective studies, the presented method promises to facilitate the identification of high-risk polyps as an automated second reader.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-08532-2.

Funding Open Access funding enabled and organized by Projekt DEAL. This study has received funding by FöFoLe, Medizinische Fakultät, Ludwig-Maximilians-Universität München, Germany (PI: Sergio Grosu).

Other financial disclosures: Benjamin M. Yeh: Consultant for Philips Healthcare, General Electric Healthcare, and Canon Medical Systems; Grants from Philips Healthcare, General Electric Healthcare. Speaker for General Electric Healthcare, Philips Healthcare, Canon Medical Systems. Shareholder, Nextrast, Inc. Royalties from UCSF patents and Oxford University Press.

Declarations

Guarantor The scientific guarantor of this publication is Michael Ingrisch.

Conflict of interest The authors declare no competing interests.

Statistics and biometry The authors Philipp Wesp, Balthasar Schachtner, and Michael Ingrisch have significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Some study subjects or cohorts have been previously reported in Grosu S, Wesp P, Graser A, et al (2021) Machine Learning-based Differentiation of Benign and Premalignant Colorectal Polyps Detected with CT Colonography in an Asymptomatic Screening Population: A Proof-of-Concept Study. Radiology 202363. https://doi.org/10.1148/radiol.2021202363.

Methodology
- retrospective
- experimental
- performed at one institution

References

1. Siegel RL, Miller KD, Jemal A (2020) Cancer statistics, 2020. CA Cancer J Clin 70:7–30. https://doi.org/10.3322/caac.21590
2. Zauber AG, Winawer SJ, O'Brien MJ et al (2012) Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med 366:687–696. https://doi.org/10.1056/NEJMoa1100370
3. Mandel JS, Bond JH, Church TR et al (1993) Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. N Engl J Med 328:1365–1371. https://doi.org/10.1056/NEJM199305133281901
4. Winawer SJ, Zauber AG, Ho MN et al (1993) Prevention of colorectal cancer by colonoscopic polypectomy. The National Polyp Study Workgroup. N Engl J Med 329:1977–1981. https://doi.org/10.1056/NEJM199312303292701
5. Kumar V, Abbas AK, Aster JC, Robbins SL (2013) Robbins basic pathology. Elsevier/Saunders, Philadelphia
6. Brenner H, Stock C, Hoffmeister M (2014) Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. BMJ 348:g2467
7. Brenner H, Altenhofen L, Kretschmann J et al (2015) Trends in adenoma detection rates during the first 10 years of the German screening colonoscopy program. Gastroenterology 149:356-366.e1. https://doi.org/10.1053/j.gastro.2015.04.012
8. Graser A, Stieber P, Nagel D et al (2009) Comparison of CT colonography, colonoscopy, sigmoidoscopy and faecal occult blood tests for the detection of advanced adenoma in an average risk population. Gut 58:241–248. https://doi.org/10.1136/gut.2008.156448
9. Kim DH, Pickhardt PJ, Taylor AJ et al (2007) CT colonography versus colonoscopy for the detection of advanced neoplasia. N Engl J Med 357:1403–1412. https://doi.org/10.1056/NEJMoa070543
10. Atkin W, Dadswell E, Wooldrage K et al (2013) Computed tomographic colonography versus colonoscopy for investigation of patients with symptoms suggestive of colorectal cancer (SIGGAR): a multicentre randomised trial. Lancet 381:1194–1202. https://doi.org/10.1016/S0140-6736(12)62186-2
11. Halligan S, Mallett S, Altman DG et al (2011) Incremental benefit of computer-aided detection when used as a second and concurrent reader of CT colonographic data: multiobserver study. Radiology 258:469–476. https://doi.org/10.1148/radiol.10100354
12. Dachman AH, Obuchowski NA, Hoffmeister JW et al (2010) Effect of computer-aided detection for CT colonography in a multireader, multicase trial. Radiology 256:827–835. https://doi.org/10.1148/radiol.10091890
13. Spada C, Hassan C, Bellini D et al (2021) Imaging alternatives to colonoscopy: CT colonography and colon capsule. European Society of Gastrointestinal Endoscopy (ESGE) and European Society of Gastrointestinal and Abdominal Radiology (ESGAR) Guideline - Update 2020. Eur Radiol 31:2967–2982. https://doi.org/10.1007/s00330-020-07413-4
14. Rex DK, Boland CR, Dominitz JA et al (2017) Colorectal cancer screening: recommendations for physicians and patients from the U.S. Multi-Society Task Force on Colorectal Cancer. Gastroenterology 153:307–323. https://doi.org/10.1053/j.gastro.2017.05.013

15. Song B, Zhang G, Lu H et al (2014) Volumetric texture features from higher-order images for diagnosis of colon lesions via CT colonography. Int J Comput Assist Radiol Surg 9:1021–1031. https://doi.org/10.1007/s11548-014-0991-2
16. Grosu S, Wesp P, Graser A et al (2021) Machine learning-based differentiation of benign and premalignant colorectal polyps detected with CT colonography in an asymptomatic screening population: a proof-of-concept study. Radiology 202363. https://doi.org/10.1148/radiol.2021202363
17. Shend D, Wu G, Suk H-I (2017) Deep learning in medical image analysis. Annu Rev Biomed Eng 19:221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442
18. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-CAM++: improved visual explanations for deep convolutional networks. IEEE Winter Conf Appl Comput Vis WACV 2018:839–847. https://doi.org/10.1109/WACV.2018.00097
19. Smith K, Clark K, Bennett W et al (2015) Data from CT_COLONOGRAPHY. Cancer Imaging Arch. https://doi.org/10.7937/K9/TCIA.2015.NWTESAY1
20. Johnson CD, Chen MH, Toledano AY et al (2008) Accuracy of CT colonography for detection of large adenomas and cancers. N Engl J Med 359:1207–1217. https://doi.org/10.1056/NEJMoa0800996
21. Clark K, Vendt B, Smith K et al (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26:1045–1057. https://doi.org/10.1007/s10278-013-9622-7
22. Nolden M, Zelzer S, Seitel A et al (2013) The Medical Imaging Interaction Toolkit: challenges and advances: 10 years of open-source development. Int J Comput Assist Radiol Surg 8:607–620. https://doi.org/10.1007/s11548-013-0840-8
23. LeCun Y, Boser B, Denker JS et al (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1:541–551. https://doi.org/10.1162/neco.1989.1.4.541
24. Chollet, Francois (2015) Keras
25. Abadi M, Barham P, Chen J, et al (2015) TensorFlow: a system for large-scale machine learning. 21. https://doi.org/10.5281/zenodo.4724125
26. Tan J, Gao Y, Liang Z et al (2019) 3D-GLCM CNN: A 3-dimensional gray-level co-occurrence matrix based CNN model for polyp classification via CT colonography. IEEE Trans Med Imaging. https://doi.org/10.1109/tmi.2019.2963177
27. Biau G, Scornet E (2016) A random forest guided tour TEST 25:197–227. https://doi.org/10.1007/s11749-016-0481-7
28. van der Meulen MP, Lansdorp-Vogelaar I, Goede SL et al (2018) Colorectal cancer: cost-effectiveness of colonoscopy versus CT colonography screening with participation rates and costs. Radiology 287:901–911. https://doi.org/10.1148/radiol.2017162359
29. Stoop EM, de Haan MC, de Wijkerslooth TR et al (2012) Participation and yield of colonoscopy versus non-cathartic CT colonography in population-based screening for colorectal cancer: a randomised controlled trial. Lancet Oncol 13:55–64. https://doi.org/10.1016/S1470-2045(11)70283-2
30. Pickhardt PJ, Hassan C, Laghi A et al (2008) Small and diminutive polyps detected at screening CT colonography: a decision analysis for referral to colonoscopy. AJR Am J Roentgenol 190:136–144. https://doi.org/10.2214/AJR.07.2646
31. Kim DH, Matkowskyj KA, Lubner MG et al (2016) Serrated polyps at CT colonography: prevalence and characteristics of the serrated polyp spectrum. Radiology 280:455–463. https://doi.org/10.1148/radiol.2016151608

6 | Publication III

International Journal of Legal Medicine (2023) 137:733–742
https://doi.org/10.1007/s00414-023-02958-7

ORIGINAL ARTICLE

Automated localization of the medial clavicular epiphyseal cartilages using an object detection network: a step towards deep learning-based forensic age assessment

Philipp Wesp[1] · Bastian Oliver Sabel[1] · Andreas Mittermeier[1] · Anna Theresa Stüber[1] · Katharina Jeblick[1,2] · Patrick Schinke[3] · Marc Mühlmann[1] · Florian Fischer[4] · Randolph Penning[4] · Jens Ricke[1] · Michael Ingrisch[1] · Balthasar Maria Schachtner[1,2]

Received: 11 October 2022 / Accepted: 24 January 2023 / Published online: 2 February 2023

Abstract

Background Deep learning is a promising technique to improve radiological age assessment. However, expensive manual annotation by experts poses a bottleneck for creating large datasets to appropriately train deep neural networks. We propose an object detection approach to automatically annotate the medial clavicular epiphyseal cartilages in computed tomography (CT) scans.

Methods The sternoclavicular joints were selected as structure-of-interest (SOI) in chest CT scans and served as an easy-to-identify proxy for the actual medial clavicular epiphyseal cartilages. CT slices containing the SOI were manually annotated with bounding boxes around the SOI. All slices in the training set were used to train the object detection network RetinaNet. Afterwards, the network was applied individually to all slices of the test scans for SOI detection. Bounding box and slice position of the detection with the highest classification score were used as the location estimate for the medial clavicular epiphyseal cartilages inside the CT scan.

Results From 100 CT scans of 82 patients, 29,656 slices were used for training and 30,846 slices from 110 CT scans of 110 different patients for testing the object detection network. The location estimate from the deep learning approach for the SOI was in a correct slice in 97/110 (88%), misplaced by one slice in 5/110 (5%), and missing in 8/110 (7%) test scans. No estimate was misplaced by more than one slice.

Conclusions We demonstrated a robust automated approach for annotating the medial clavicular epiphyseal cartilages. This enables training and testing of deep neural networks for age assessment.

Keywords Anatomic landmark detection · Deep learning · Object detection · Medial clavicular epiphyseal cartilages · Age assessment

Background

Age is an essential part of a person's identity, especially for children. By definition of the UN Convention on the Rights of the Child (CRC, Article 1) [1] and the EU acquis (Directive 2013/33/EU, Article 2(d)) [2], a child is any person below the age of 18. When the age is known, it rules the relationship between a person and the state. Changes in age can trigger the acquisition of rights and obligations in different aspects such as emancipation, employment, criminal responsibility, sexual relation, and consent for marriage or military service [3]. Because of the importance of age, the CRC lists certain key obligations for states and authorities regarding age that include registration of the child after

✉ Philipp Wesp
philipp.wesp@med.uni-muenchen.de

1 Department of Radiology, University Hospital, LMU Munich, Marchioninistraße 15, 81377 Munich, Germany

2 Comprehensive Pneumology Center (CPC-M), Member of the German Center for Lung Research (DZL), Max-Lebsche-Platz 31, 81377 Munich, Germany

3 Institute of Informatics, LMU Munich, Oettingenstraße 67, 80538 Munich, Germany

4 Institute of Forensic Medicine, LMU Munich, Nußbaumstraße 26, 80336 Munich, Germany

birth, respecting the right of the child to preserve his or her identity, and speedily re-establish his or her identity in case that some or all elements of the child's identity have been deprived [3]. Following these obligations, a state may need to assess the age of the person to determine whether the person is an adult or a child when the age is unknown. In that case, the European Union Agency for Asylum (EUAA) recommends that the least intrusive method is selected following a gradual implementation and that the most accurate method is selected and margin of error is documented [3].

Radiological examinations of the carpal bones, the molars or the clavicles play an important role in assessing the chronological age of living individuals [4]. For the clavicles, the ossification status of the medial clavicular epiphyseal cartilages is of particular interest. As the last maturing bone structure in the body, it allows age assessment not only for minors, but also for young adults [5]. However, current standard methods for age assessment suffer from low accuracy, intra- and inter-reader variability, and low diversity within the study populations [4, 6, 7].

A promising approach for accurate and automated age assessment is deep learning. Deep learning has been applied successfully to a wide range of computer vision tasks in medical imaging in the past [8–10]. A deep neural network trained to map an image of the medial clavicular epiphyseal cartilages to an individual's age may yield more accurate age assessment results compared to current approaches [8, 11, 12]. The data required to train a deep network for age assessment, i.e., medical images including clavicles and sternum, as well as information about the age of the corresponding individuals, is abundant in many hospitals and also easily accessible. However, for efficient and successful training, it is advisable to first localize the medial clavicular epiphyseal cartilages within the medical images. The training process for diagnostic computer vision networks benefits from inputs that are cropped to the image region containing information relevant for solving the problem [13]. This cropping step usually requires manual expert annotations, which are time-consuming and expensive [14].

Therefore, the aim of this study was to develop and to evaluate an automated approach to localize the medial clavicular epiphyseal cartilages in CT scans, using deep learning–based object detection. This automated localization can be used to create large datasets, which are necessary to appropriately train and evaluate a deep neural network for age assessment [15].

Methods

We propose to use the state-of-the-art object detection network RetinaNet [16] for the automated localization of the medial clavicular epiphyseal cartilages in CT scans. First, a trained instance of the two-dimensional RetinaNet was applied to each axial slice in a scan in order to detect a proxy structure for the medial clavicular epiphyseal cartilages (Fig. 1). In case of a detection, the RetinaNet predicted a bounding box, as well as a class, and provided a classification score. Multiple detections in different slices or within the same slice were possible. The center of the bounding box associated with the highest of all classification scores was entitled as the location estimate of the medial clavicular epiphyseal cartilages in the CT scan. The entire workflow is illustrated in Fig. 2.

In the following, this section will describe (a) the retrospective collection of the data; (b) the manual data annotation; (c) the splitting of the data into training, validation, and test set; (d) the object detection network RetinaNet; (e) the training and evaluation of the RetinaNet; and (f) how we used the RetinaNet to estimate the location of the medial clavicular epiphyseal cartilages in a scan.

Retrospective data collection

This study was approved by the institutional review board, and the requirement for written informed consent was waived. CT scans of the upper body were identified retrospectively in the picture archiving and communication system (PACS). The scans were originally acquired during the clinical routine for all purposes in the period 2017–2020. The patients' age at examination was in the range of 15 to 25 years; age was measured as the time

Fig. 1 The structure-of-interest (SOI), defined as the sternoclavicular joints, together with their contributing portions of the sternum and the medial clavicles

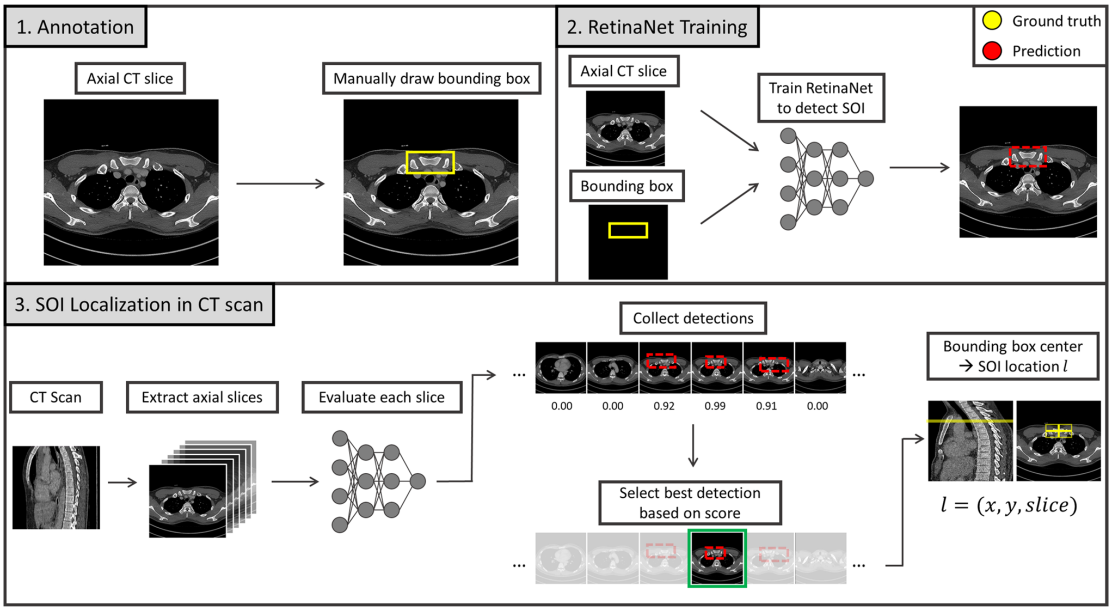

Fig. 2 Schematic workflow diagram of the proposed medial clavicular epiphyseal cartilage localization. 1. Annotation: CT images are manually annotated with two-dimensional ground-truth bounding boxes in axial slices around the structure-of-interest (SOI). The SOI is an easy-to-identify proxy structure for the actual medial clavicular epiphyseal cartilage. 2. RetinaNet Training: A RetinaNet is trained to detect the SOI in axial slices and predict bounding boxes. 3: Localization in CT scan: The SOI can be localized in an unknown CT scan of the upper body. For this purpose, the trained RetinaNet is applied to each slice in a CT scan and all positive detections are collected. Afterwards, the center of the bounding box which corresponds to the best detection (highest classification score) is used as the predicted location for the SOI

difference in days between documented date of birth and date of CT examination. On the one hand, this range covers a broad spectrum of developmental stages of the medial clavicular epiphyseal cartilages [17]; on the other hand, it includes ages which have high legal relevance in most countries, e.g., R18 and 21 [4]. Detailed inclusion and exclusion criteria for CT scans are listed in the "Appendix."

Three preprocessing steps were applied to the collected scans. First, image voxel values were limited to the range of − 200 to 600 Hounsfield units (HUs). This value range was derived heuristically, with the intent to remove information from the image that we considered less relevant for the detection of the proxy structure for the medial clavicular epiphyseal cartilages. This signal intensity restriction was supposed to guide the network to focus on a balanced mix of bone and clavicles surrounding soft tissue. Second, axial slices have been resized to 512×512 pixels to match the input size that is expected by the RetinaNet. Finally, pixel values in each axial slice were linearly scaled into the value range 0.0 to 1.0 for network training.

Manual data annotation

The structure-of-interest (SOI) in this study was defined as the sternoclavicular joints, together with their contributing portions of the sternum and the medial clavicles (Fig. 1). The SOI served as an easy-to-identify proxy for the actual medial clavicular epiphyseal cartilages and was to be the structure detected by the RetinaNet.

All axial slices from the collected CT scans were manually annotated with ground-truth target labels for the RetinaNet. These target labels have two components: first, a bounding box which located the object, represented by 4 parameters—(a) x position, (b) y position, (c) width, and (d) height—and second, a class label which classified the object. If a slice included the SOI, a target label was created by manually drawing a bounding box around all visible portions of the sternum and the medial clavicles contributing to the sternoclavicular joints and setting the class label of the object to "sternum" (Fig. 3). Depending on the patient and scanning protocol, in particular the slice thickness, multiple consecutive axial slices contained the SOI and were annotated with bounding boxes and class labels accordingly.

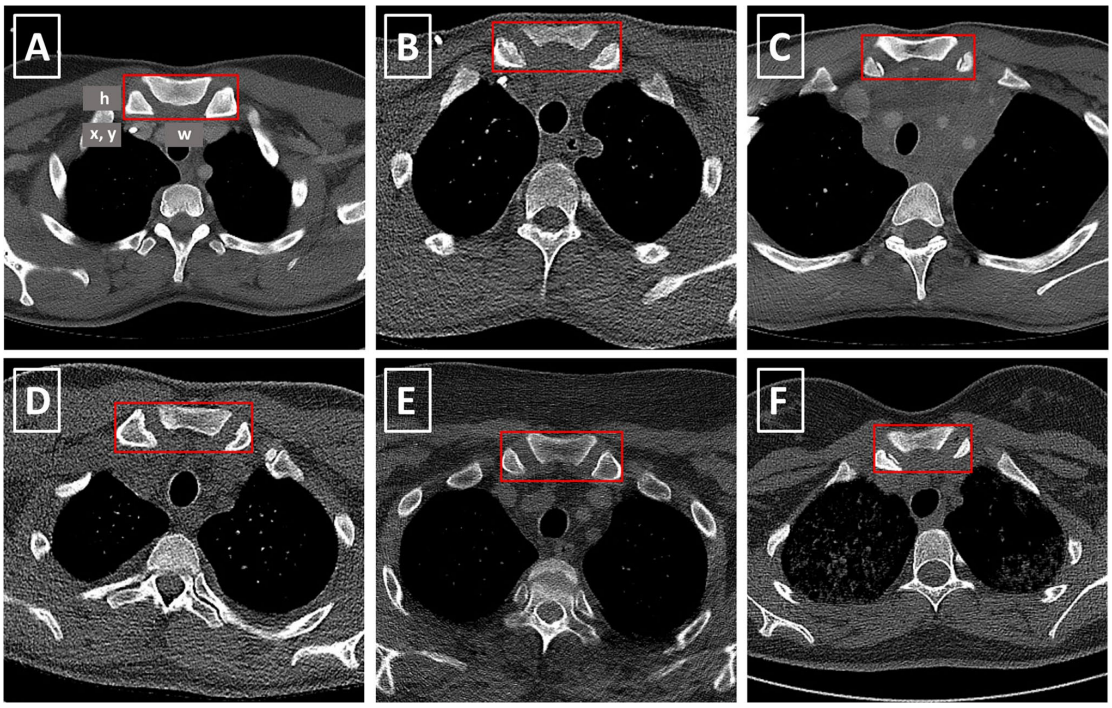

Fig. 3 (**A**–**F**) Bounding boxes around the SOI in different CT scans after preprocessing. The SOI is defined as the sternoclavicular joints, together with their contributing portions of the sternum and the medial clavicles. In addition, (**A**) illustrates the 4 bounding box location parameters x, y, width (w), and height (h)

Training, validation and test set

The network was evaluated using the three-way holdout method. To this end, the $n = 222$ collected and annotated CT scans were split into three sets: (a) training, (b) validation, and (c) testing. First, a test set consisting of 110 scans from 110 patients was randomly selected from the whole study dataset, so that it contained $n = 10$ patients and scans per age in years (ages = 15, 16, …, 25). All remaining $n = 112$ scans that have not been selected for the test set were split according to a 90/10 ratio into a training set of $n = 100$ (= Floor[0.9×112]) scans and a validation set of $n = 12$ (= 112—100) scans. The test set was used only for the evaluation of the RetinaNet, once the training was completed. The training set was used to train the RetinaNet, while the validation set was used to monitor the training process. No resampling strategy, such as cross-validation, was applied.

Object detection network

The automated localization of the SOI proposed in this study was based on a PyTorch implementation [18, 19] of the object detection network RetinaNet [16]. This RetinaNet had a ResNet18 [20] backbone, which was provided by PyTorch as an off-the-shelf network and had been pre-trained on the public dataset ImageNet [21]. An important component of the RetinaNet implementation was the Focal Loss [16], which addressed heavy class imbalance for one-stage object detectors like the RetinaNet. This was useful, as the majority of an upper body CT scan does not cover the SOI.

As input, the network expects an image size of 512×512 pixels, being a two-dimensional axial slice from a preprocessed CT scan. In case of a detection, the network returns three outputs: (a) a bounding box prediction which locates the detected object, (b) a class prediction which classifies the detected object, and (c) a classification score between 0.0 and 1.0 quantifying the confidence of the network in the predicted detection. The class prediction is trivial, as "sternum" is the only class. Higher classification scores imply increased confidence in the detection.

Object detection training and evaluation

The RetinaNet was trained for 20 epochs with examples from the training set using the Adam optimization algorithm [22]

and a base learning rate of 10^{-5} to minimize the focal loss. A learning-rate scheduler decreased the learning rate by a factor of 10 whenever the loss did not improve for 3 consecutive epochs. Data augmentation was applied: during training, we randomly flipped the image input and the bounding box along the same randomly chosen axis. Training progress was monitored by evaluating the loss of the validation set.

After training, the RetinaNet was applied to and evaluated on the test set. The predicted bounding boxes and class labels were compared to the manually annotated ground-truth targets to identify positive and negative detections—the term negative detection can be used interchangeably with no detection. A classification score ≥ 0.05 is considered a positive detection. The detection is true positive, if the intersection over union (IoU) for the areas of the predicted bounding box A_{pred} and the ground-truth bounding box A_{true} is > 0.5 and the predicted class is the ground-truth class. Otherwise, the detection is considered false positive.

$$IoU\left(A_{pred}, A_{true}\right) = \frac{\left|A_{pred} \cap A_{true}\right|}{\left|A_{true} \cup A_{pred}\right|}$$

A classification score < 0.05 (value adapted from [16]) is a true-negative detection, if the image does not contain a bounding box labeled as "sternum." Otherwise, it is a false-negative detection.

Network performance was evaluated using average precision (AP) [23], a popular metric for object detection since it was applied for the PASCAL Visual Object Classes (VOC) Challenge in 2007 [24, 25]. AP is calculated as the area under the precision-recall curve from all positive and negative network detections for the test set, ranked according to classification score in descending order, where the precision p is set to the maximum precision obtained for any recall $r^{'} \geq r$ [25]:

$$AP = \sum_{n} (r_{n+1} - r_n) \cdot p_{interp}(r_{n+1})$$

$$p_{interp}\left(r_n\right) = \max p\left(r^{'}\right), \quad r^{'} : r' \geq r_n$$

Estimating the location of the SOI

The RetinaNet was trained to detect the presence of the SOI in axial CT slices. Because the SOI is a structure which typically stretches across multiple axial slices in a CT, the network may return positive detections for multiple slices. However, for data annotation purposes, we wanted the localization approach to yield a unique location estimate of the SOI for a given CT scan.

Estimating the location of the SOI included the following steps (see Fig. 2): (a) apply the RetinaNet to each slice in a given CT scan, (b) collect all positive detections, (c) select the best detection based on the classification score, and (d) select the center of the bounding box of the best detection to be the unique estimated location of the SOI. For example, when given a CT scan consisting of 300 axial slices, the RetinaNet may detect the SOI in slices 240 and 241. The detection in slice 241 may have a classification score of 0.96, while slice 240 may only have a score of 0.92. In that case, the bounding box center of the detection in slice 241 would be the unique estimated location of the SOI. In this context, a location encoded the position in three dimensions (x, y, slice): the position in the axial plane (x, y) and the number of the slice of the respective detection counting in axial direction (slice).

Location estimates were evaluated per scan. Location estimates based on true-positive detections were also true positives. Location estimates based on false-positive detections were also false positives. Location estimates for scans with no positive detection were automatically false negatives, because each scan contained the SOI. There were no true-negative location estimates. We also evaluated the Euclidean distance between the estimated location and the center of the (nearest) ground-truth bounding box in the axial plane. Additionally, we evaluated the number of slices between the estimated location and the center of the (nearest) ground-truth bounding box.

Results

Data

The retrospectively collected image data (training set, validation set, and test set) in this study comprised 63,999 two-dimensional axial slices from 222 CT scans and 202 patients (86 female (42.6%)) from age 15 to 25. In total, 872/63,999 (1.4%) slices include the SOI and were annotated with the class label "sternum" and with a ground-truth bounding box.

The total image data was divided into three sets: training set, validation set, and test set (Table 1). The test set consisted of 30,846 slices from 110 scans and 110 patients (50 female (45.5%)); 379/30,846 (1.2%) slices included the

Table 1 Training, validation, and test set composition with respect to the number of patients, CT scans, axial slices, and annotated slices

Dataset	Patients	Scans	Slices	Annotated slices
Training	82	100	29,656	434 (1.5%)
Validation	10	12	3,497	41 (1.2%)
Test	110	110	30,846	379 (1.2%)

SOI, were labeled as class "sternum," and had a ground-truth bounding box. The training set consisted of 29,656 slices from 100 scans and 82 patients (35 female (42.7%)); 434/29,656 (1.5%) slices included the SOI, were labeled as class "sternum," and had a ground-truth bounding box. The validation set consisted of 3497 slices from 12 scans and 10 patients (1 female (10.0%)); 41/3,497 (1.2%) slices included the SOI, were labeled as class "sternum," and had a ground-truth bounding box.

Object detection network

The trained RetinaNet achieved an AP of 0.82 (1.0 = perfect score) for the detection, i.e., simultaneous localization and classification, of the SOI in the two-dimensional axial CT scan slices of the test set. The average IoU of the bounding boxes predicted by the RetinaNet and the ground-truth bounding boxes was 0.74 (1.0 = identical boxes; 0.0 = no overlap between boxes).

For the 379 slices in the test set which included the SOI, the network yielded 338/379 (89.2%) true-positive detections, and 41/379 (10.8%) false-negative detections (Table 2). Examples of a true-positive detection and a false-negative detection are shown in Fig. 4. The median classification score for the 379 test slices that include the SOI was 1.00 [lower quartile (LQ) = 0.98; upper quartile (UQ) = 1.00] (1.0 = perfect score). The median IoU of the predicted bounding boxes and ground-truth bounding boxes in these slices was 0.83 [LQ = 0.76; UQ = 0.88].

For the 30,467 (= 30,846—379) slices in the test set that did not include the SOI, the network yielded 51/30,467 (0.2%) false-positive detections and 30,416/30,467 (99.8%) true-negative detections (Table 2). The median classification score for the 51 false-positive detections was 0.88 [LQ = 0.17; UQ = 1.00]. The median classification score for the 30,416 true-negative detections was 0.00 [LQ = 0.00; UQ = 0.00].

Estimating the location of the SOI

The center of the bounding box from the RetinaNet detection with the highest classification score of all detections in a given CT scan was the estimated location of the SOI for that scan. Estimated locations were compared to ground-truth locations.

In 97/110 (88%) scans of the test set, the estimated location was true positive, i.e., in a slice with a ground-truth location. In 5/110 (5%) scans of the test set, the estimated location was false positive, but in slices directly next to a slice with a ground-truth location. In 8/110 (7%) scans of the test set the location estimate was false negative, because the RetinaNet did not return a positive detection, despite the SOI being present in the scan. The classification score distribution returned by the RetinaNet for a scan of the test set and the slice of the estimated SOI location is included in Fig. 4.

For the 97 true-positive location estimates, the mean (standard deviation (SD)) distance in the axial plane between the estimated location and the true location was 6.0 (3.8) pixels. For the 5 false-positive location estimates, the mean (SD) distance in the axial plane between the estimated location and the closest true location was 7.6 (3.7) pixel. The average number of slices between a false-positive location estimate and the closest true location was 1 slice.

Table 2 Confusion matrix of RetinaNet detections in the test set. Detections in CT slices which include the SOI and have an IoU > 0.5 with the ground-truth bounding box are true positives. A CT slice which does not include the SOI and for which the RetinaNet did not yield a detection is a true negative

	Detection in slice	No detection in slice
SOI in slice	338 / 379 (89.2%)	41 / 379 (10.8%)
SOI not in slice	51 / 30,467 (0.2%)	30,416 / 30,467 (99.8%)

Discussion

We investigated a deep learning approach based on the state-of-the-art object detection network RetinaNet in order to locate the medial clavicular epiphyseal cartilages through an easy-to-identify proxy structure: the SOI. The dedicated RetinaNet trained in this study achieved an AP of 0.82 for detecting the SOI in all axial CT slices of the test set. Based on the RetinaNet detections, the location of the SOI was estimated correctly in 88% of the CT scans in the test set, the false-positive localizations (5%) being close misses and the false negatives (7%) not being harmful. These results show that the presented localization approach can be used to reliably generate large amounts of annotated data for training and evaluating a dedicated deep neural network for age-assessment, without being limited by expensive and time-consuming manual annotations through medical experts. A large dataset is necessary to train high-performing deep neural networks for any given task [15]. Using the localization approach as a foundation, deep learning–based age estimation has the potential to be more accurate than today's standard approaches [8, 11, 12]. In addition, the localization approach enables automated end-to-end age assessment without human interaction that only requires a CT scan which includes the medial clavicular epiphyseal cartilages as input.

The presented localization approach enables large annotated datasets for deep learning–based age assessment for multiple reasons. First, in the majority of the test scans (97/110 (88%)), the predicted location of the SOI was in a correct axial slice and only 6.0 pixels away from the ground-truth location on average. Even in the small number of test

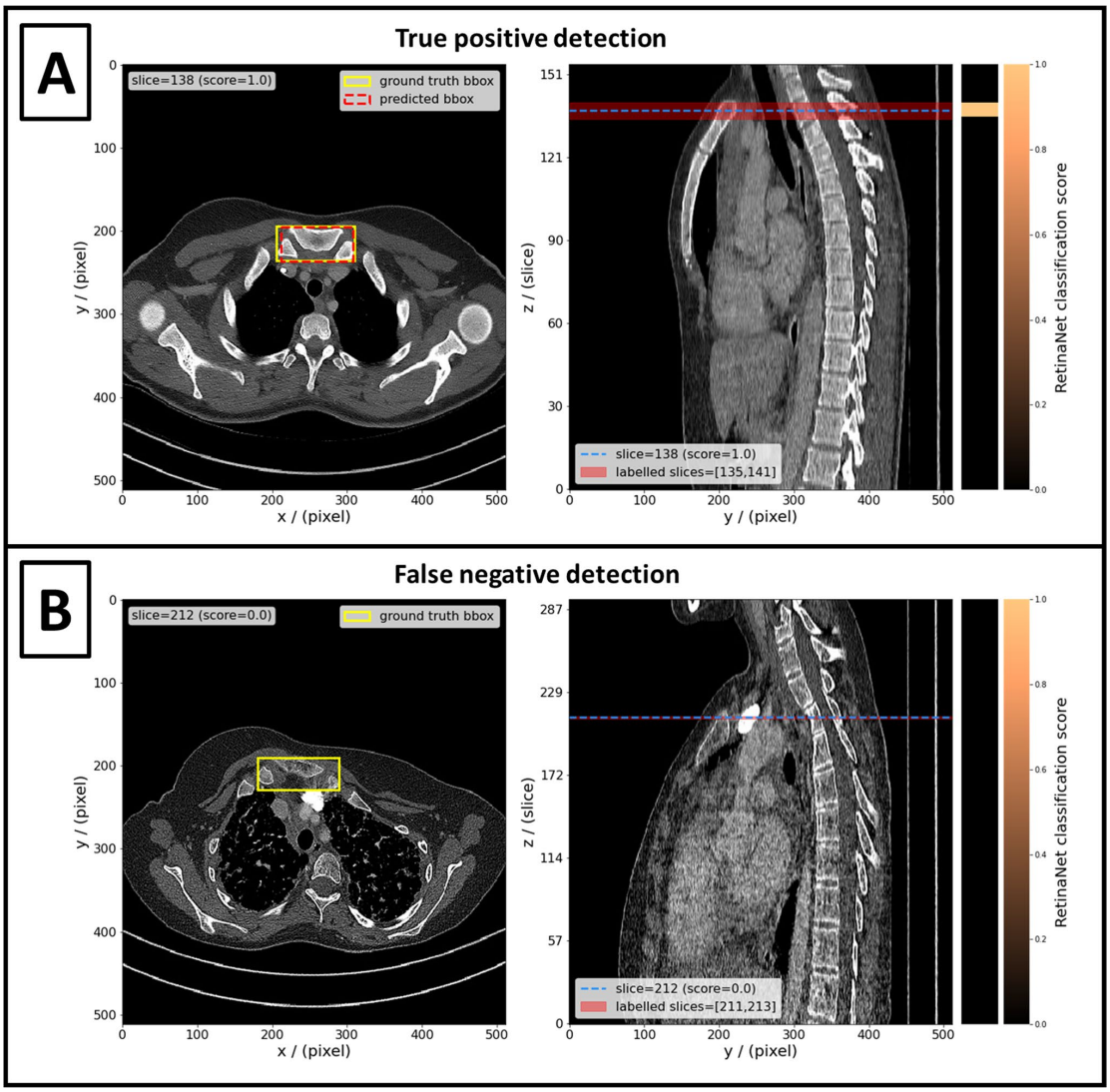

Fig. 4 The left panels show axial CT slices with ground-truth bounding boxes around the SOI (yellow boxes) and detections (if predicted by network) (red boxes). The right panels show the central sagittal slice of the respective CT. The position of the axial slice in the left panel is indicated by the dashed blue line in the right panel. The red area in the right panel indicates the positions of all axial slices which contain the SOI and have ground-truth bounding boxes annotated. The heatmaps next to the right panels show the classification score returned by the RetinaNet for each axial slice (light orange = 1.0; black = 0.0). Detections made by the RetinaNet are true positive, if the axial slice has a ground-truth bounding box (red area) and the classification score is > 0.05 (e.g., light orange). **A** Shows an example of a true-positive localization of the SOI; i.e., the highest classification score was returned for a slice which indeed contains the SOI. **B** shows an example of a false-negative localization; i.e., the RetinaNet returned only classification scores < 0.05 even though the SOI is present in one or more slices

scans with false-positive detections (5/110 (5%)), the SOI was misplaced by only one slice. The three-dimensional field of view of a deep neural network for age assessment could most likely be chosen large enough, so that the localization in these five respective scans would still be sufficient. Moreover, in all remaining test scans (8/110 (7%)) the RetinaNet did not yield a detection. Although the number of false negatives should be reduced in the future, they are unproblematic for

the generation of a dataset for deep learning–based age assessment. False negatives only reduce the number of annotated scans that can be generated from a given amount of unlabeled CT scans. As long as the required number of cases remains feasible, negative detections could also trigger the need for manual annotation of the medial clavicular epiphyseal cartilages through medical experts. This way, every available CT scan including the medial clavicular epiphyseal cartilages may be used as training data for a deep age-assessment network.

To the best of our knowledge, there exists no comparable anatomical landmarking approach for the purpose of locating the medial clavicular epiphyseal cartilages, the sternum, or the clavicles in CT scans. However, anatomic landmarking in medical images is an active research field, and there are a variety of studies which apply deep learning to locate different anatomical structures for distinct purposes [26, 27]. A particular application for anatomic landmarking in medical images, which also shares some conceptual overlap with our study, is the detection of bone fractures. In these studies, deep object detection networks could successfully be trained to detect cracks in bone tissue and to locate fractures in hand and chest radiographs or chest CT scans [27]. Among other areas, one particular network was able to draw a bounding box around the clavicles in radiographic images in case of a present fracture [26].

There are limitations within this study. First, the number of patients ($n = 202$) and CT scans ($n = 222$) was small, because manual ground-truth annotations ($n = 872$ bounding boxes) were time-consuming and data acquisition through the PACS laborious. However, the dataset was deemed large enough to perform this first study. Second, the training set included 100 CT scans from 82 patients; the validation set included 12 CT scans from 10 patients, which means that patient doublets were presented to the RetinaNet in each training epoch. The prevention of doublets is generally considered a quality standard regarding the reference population. The test set used for evaluation did not contain doublets. Third, the dataset is limited to CT images, and no statement about the performance of the automated localization approach for MRI images can be made. CT images were used because CT is the state-of-the-art for forensic questions as it is widely available, quick, cheap, and robust. However, MRI is more desirable for acquiring images in healthy individuals compared to CT, because it spares the individuals from harmful ionizing radiation. But, we believe that the approach can be translated to MRI images in the future. Next, the approach does not differentiate between the left and right sternoclavicular joint and instead locates a proxy structure which includes both joints. As the differentiation of left and right clavicles is crucial in forensic age estimation, expanding the capabilities of the localization in that regard would be an interesting future step. Additionally, the CT scans in this study were originally acquired during the clinical routine for all purposes. Because we were not able to analyze these purposes, there could be a bias in our dataset. Also, the observed false-negative detections may occur systematically and excluding them from a dataset for deep learning–based age assessment could introduce a bias. Furthermore, the thresholds for limiting HU values were derived heuristically and the potential effect of different thresholds on localization performance was not measured. Finally, we did not investigate three-dimensional object detection, even though it would have been natural to the problem of locating the medial clavicular epiphyseal cartilages in a CT scan. However, compared to 3D object detection, 2D object detection is much more common [28] and has a lot of benefits: (a) for the same amount of CT scans, more 2D slices than 3D scans that can be used as training examples, (b) 2D inputs are smaller and allow using smaller networks with fewer parameters, and (c) a wide range of high-performing pretrained models is available for 2D inputs.

Conclusions

In summary, we demonstrated a robust deep learning–based localization of an anatomical proxy structure to automate the localization of the medial clavicular epiphyseal cartilages. This enables deep learning–based age estimation based on the ossification of the medial clavicular epiphyseal cartilages which might outperform today's standard methods. The presented localization approach addresses a specific case of a much wider problem concerning machine learning in medicine: human annotations are costly and difficult to acquire, while the lack of annotations poses an enormous bottleneck for machine learning performance [14, 29].

Appendix

CT inclusion and exclusion criteria

The first two inclusion criteria were applied to identify clinical studies in the PACS. Inclusion criteria (PACS):

- Study includes chest CT
- Patient was between 15 and 25 years of age at the time of chest CT acquisition

The remaining 9 inclusion and exclusion criteria are listed below and were applied by inspecting the respective DICOM tags of the chest CT scans. The criteria were applied to each chest CT scan of the identified studies. The list of accepted reconstruction kernels was handcrafted based on all hard kernels we encountered during the retrospective data collection. We did not explicitly filter CT scans based on slice

thickness. However, we used slice thickness to include only one chest CT scan per study and thereby keep the number of patients' duplicates small. Specifically, we selected the chest CT scan with the thinnest slice thickness and discarded all other scans, in case the PACS query yielded multiple eligible chest CT scans for the same study. Nevertheless, patient duplicates exist in the dataset, because some patients were subject to multiple studies. Inclusion criteria (DICOM):

- *Modality* is "CT"
- *ImageOrientationPatient* is [1,0,0,0,1,0]
- *ContrastBolusAgent* is None or ""
- *ConvolutionKernel* is "LUNG," "BL57f/3," "BL57d/3," "I70f/3," "I70f\3," ["I70f," "3"], "I70f/2," "I70f\2" or ["I70f", "2"]

Exclusion criteria:

- "patient protocol" in *SeriesDescription.lower()*
- "topogram" in *SeriesDescription.lower()*
- "mpr" in *SeriesDescription.lower()*
- "SPO" in *SeriesDescription*
- "mip" in *SeriesDescription.lower()*

Slice thickness

The median slice thickness of the 222 CT scans in this study was 2.0 mm, the minimum 0.625 mm, the maximum 3.0 mm, the lower quartile 1.0 mm, and the upper quartile 2.5 mm. The distribution of the 222 slice thickness values is shown in Fig. 5.

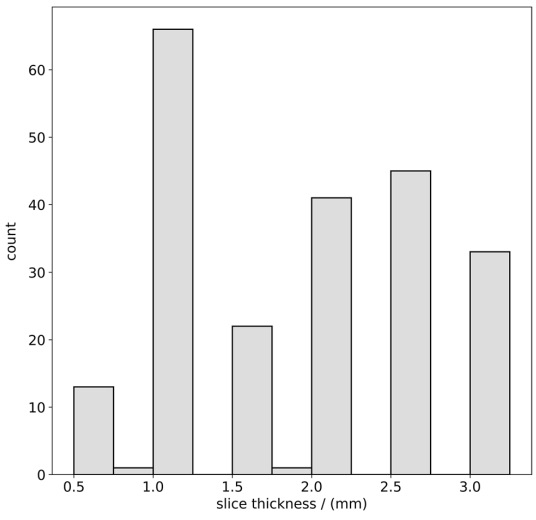

Fig. 5 Slice thickness distribution of the 222 chest CT scans

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The datasets generated and analyzed during the current study are not publicly available due to them containing information that could compromise research participant privacy, but are available from the corresponding author on reasonable request and with permission of the institutional review board (Ethics Committee, Medical Faculty, LMU Munich). Custom code described in the manuscript is available in the GitHub repository: https://github.com/pwesp/automated-clavicular-epiphysis-localization.

Declarations

Ethics approval This retrospective study was approved by the institutional review board (Ethics Committee, Medical Faculty, LMU Munich). All methods were carried out in accordance with the Declaration of Helsinki.

Competing interests The authors declare no competing interests.

Informed consent The requirement for written informed consent was waived by the institutional review board (Ethics Committee, Medical Faculty, LMU Munich).

References

1. United Nations (1989) The convention on the rights of the child. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child. Accessed 31 Jan 2023
2. The European Parliament and The Council Of The European Union (2013) Directive 2013/33/EU of the European Parliament and of the Council of 26 June 2013 laying down standards for the reception of applicants for international protection (recast). https://eur-lex.europa.eu/eli/dir/2013/33/oj. Accessed 31 Jan 2023
3. European Asylum Support Office (2018) EASO Practical guide on age assessment. Publications Office. https://doi.org/10.2847/292263
4. Schmeling A, Dettmeyer R, Rudolf E et al (2016) Forensic age estimation: methods, certainty, and the law. Dtsch Arzteblatt Int 113:44–50. https://doi.org/10.3238/arztebl.2016.0044
5. Schmeling A, Schulz R, Reisinger W et al (2004) Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med 118:5–8. https://doi.org/10.1007/s00414-003-0404-5
6. Kellinghaus M, Schulz R, Vieth V et al (2010) Forensic age estimation in living subjects based on the ossification status of the medial clavicular epiphysis as revealed by thin-slice multidetector

computed tomography. Int J Legal Med 124:149–154. https://doi.org/10.1007/s00414-009-0398-8
7. Kellinghaus M, Schulz R, Vieth V et al (2010) Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans. Int J Legal Med 124:321–325. https://doi.org/10.1007/s00414-010-0448-2
8. Halabi SS, Prevedello LM, Kalpathy-Cramer J et al (2019) The RSNA pediatric bone age machine learning challenge. Radiology 290:498–503. https://doi.org/10.1148/radiol.2018180736
9. Sahiner B, Pezeshk A, Hadjiiski LM et al (2019) Deep learning in medical imaging and radiation therapy. Med Phys 46:e1–e36. https://doi.org/10.1002/mp.13264
10. Shen D, Wu G, Suk H-I (2017) Deep learning in medical image analysis. Annu Rev Biomed Eng 19:221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442
11. Obermeyer Z, Emanuel EJ (2016) Predicting the future — big data, machine learning, and clinical medicine. N Engl J Med 375:1216–1219. https://doi.org/10.1056/NEJMp1606181
12. Wittschieber D, Schulz R, Vieth V et al (2014) The value of sub-stages and thin slices for the assessment of the medial clavicular epiphysis: a prospective multi-center CT study. Forensic Sci Med Pathol 10:163–169. https://doi.org/10.1007/s12024-013-9511-x
13. Diligenti M, Roychowdhury S, Gori M (2017) Integrating prior knowledge into deep learning. 16th IEEE International Conference on Machine Learning and Applications (ICMLA). https://doi.org/10.1109/ICMLA.2017.00-37
14. Willemink MJ, Koszek WA, Hardell C et al (2020) Preparing medical imaging data for machine learning. Radiology 295:4–15
15. Chartrand G, Cheng PM, Vorontsov E et al (2017) Deep learning: a primer for radiologists. Radiographics 37:2113–2131. https://doi.org/10.1148/rg.2017170077
16. Lin TY, Goyal P, Girshick R et al (2020) Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell 42:318–327. https://doi.org/10.1109/TPAMI.2018.2858826
17. Cunningham C, Scheuer L, Black S (2016) Developmental juvenile osteology. Academic Press, Cambridge, MA, USA
18. Paszke A, Gross S, Massa F et al (2019) PyTorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems. Curran Associates, Red Hook, NY, USA. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf. Accessed 31 Jan 2023
19. (2018) Pytorch-retinanet. https://github.com/yhenon/pytorch-retinanet. Accessed 31 Jan 2023
20. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp 770–778. https://doi.org/10.1109/CVPR.2016.90
21. Deng J, Dong W, Socher R et al (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
22. Kingma DP, Ba JL (2015) Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR). https://arxiv.org/pdf/1412.6980.pdf. Accessed 31 Jan 2023
23. Salton G, McGill MJ (1986) Introduction to modern information retrieval. McGraw-Hill, New York City, NY, USA
24. Everingham M, Van Gool L, Williams CKI et al (2010) The pascal visual object classes (VOC) challenge. Int J Comput Vis 88:303–338. https://doi.org/10.1007/s11263-009-0275-4
25. Everingham M, Winn J (2010) The PASCAL visual object classes challenge 2010 (VOC2010) Development Kit. http://host.robots.ox.ac.uk/pascal/VOC/voc2010/devkit_doc_08-May-2010.pdf. Accessed 31 Jan 2023
26. Ma Y, Luo Y (2021) Bone fracture detection through the two-stage system of Crack-Sensitive Convolutional Neural Network. Inform Med Unlocked 22:100452. https://doi.org/10.1016/j.imu.2020.100452
27. Lindsey R, Daluiski A, Chopra S et al (2018) Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 115:11591–11596. https://doi.org/10.1073/pnas.1806905115
28. Jiao L, Zhang F, Liu F et al (2019) A survey of deep learning-based object detection. IEEE Access 7:128837–128868. https://doi.org/10.1109/ACCESS.2019.2939201
29. Rajkomar A, Dean J, Kohane I (2019) Machine Learning in Medicine. N Engl J Med 380:1347–1358. https://doi.org/10.1056/nejmra1814259

7 | Conclusion

This work demonstrates proof-of-concept machine learning approaches in two areas of radiology: colorectal cancer screening with computed tomography (CT) colonography and radiological age assessment based on clavicle ossification in CT. Machine learning allowed to overcome challenges where the capabilities of conventional imaging diagnostics reach their limits.

CT colonography This work shows that classical machine learning and deep learning enable the non-invasive classification of CT colonography–detected colorectal polyps. The random forest model can differentiate premalignant and benign colorectal polyps using radiomic features extracted from CT colonography scans [12]. These features are calculated based on polyp segmentation masks that were manually annotated by experienced radiologists. The model performance was validated in an external multicenter test set with a histopathologic reference standard. A feature importance analysis revealed that only one of the 10 most relevant radiomic features for decision-making is a size-measuring feature, while the other nine features characterize texture or first-order histogram statistics. This is particularly interesting because current guidelines suggest using size as a surrogate indicator for malignancy [20, 21].

Proceeding from these results, this work also explored a convolutional neural network (CNN) for solving the same task. The model was trained on the same training data and evaluated with the same external multicenter test set as the random forest. The results show that the CNN enables polyp differentiation even when the model is not provided with polyp segmentation masks [13]. This is a valuable advantage compared to the random forest approach because manual segmentation by experts is a barrier to the potential integration of machine learning based CT colonography analysis into the clinical routine and prevents fully automated polyp classification. Furthermore, deep learning enables visual interpretability by highlighting image regions that were potentially important for the model's predictions of lesion character.

However, CNN performance was worse compared to the random forest, mainly when no segmentation masks were provided. This is likely to be caused by the size of the training dataset. Although the dataset is of high quality and has ground truth la-

bels according to a histopathological reference standard, the number of examples to learn from is small for deep learning. Expanding the training dataset in the future is difficult because CT colonography is not yet a frequent routine examination at LMU University Hospital and not every polypectomy with subsequent histopathological analysis is performed in-house. One approach for a larger dataset that might lead to increased deep learning performance would be a retrospective national or international, multicentric study that combines the CT colonography data from different sites. An expensive alternative would be a large, prospective study over to course of several years at LMU University Hospital, similar to the study performed by Graser et al. [19] that gathered the current dataset, but with more participants.

Overall, the identification of high-risk colorectal polyps using machine learning enables individual risk stratification and therapy guidance for CT colonography examinations. Additionally, the CNN results are an encouraging step towards using machine learning as an automated second reader.

Radiological age assessment This work displays continuous predictions of chronological age based on clavicle ossification in CT using a combination of a deep learning object detection model and a CNN that outperforms an established standard method on average. The object detection model reliably locates an anatomical proxy structure for the medial clavicular epiphyseal cartilages in thoracic CT scans, which eliminates the need for expensive manual clavicle localization by human experts [14]. Thereby, the object detection model enabled the automated creation of a large dataset of CT images cropped to the anatomical region containing the ossification status of the clavicles. In deep learning, it is advisable to use training examples that contain relevant information and are tailored to the problem that should be solved [40]. This large dataset was the foundation for developing a deep learning model for radiological age assessment based on clavicle ossification. Thus, object detection solves a specific instance of a much broader problem of machine learning in radiology: the acquisition of human annotations is costly and challenging, and the lack of annotations is a big bottleneck for machine learning performance.

Using the large dataset, a CNN was successfully trained in this work to map thoracic CT scans cropped around the medial clavicular epiphyseal cartilages to chronological age [15]. The automated and continuous age predictions with deep learning are more accurate on average compared to an optimistic performance estimate for the well-recognized classical age assessment method of Kellinghaus et al. [37, 38]. However, deep learning predictions also show a higher variance, and the highest absolute errors were observed for deep learning when comparing both methods. In cases where the deep learning prediction was less accurate than the human reader performance estimate, poor deep learning performance could partially be attributed to norm-variants or pathologic disorders of the clavicles. Typically, such physiological abnormalities would be exclusion criteria for radiological age assessment with reference study methods [108, 109]. Also, the highest deep learning prediction errors can be avoided by

abstaining from predictions when the model shows high predictive uncertainty.

These issues could potentially be addressed in the future by expanding the training dataset and implementing additional exclusion criteria to avoid physiological abnormalities in the training examples. Larger datasets might also facilitate the successful training of more complex deep learning approaches, e.g. vision transformers, that potentially yield higher accuracy while minimizing extreme errors. Also, it is worth noting, that the human reader performance estimate assumes a best-case scenario that favors human performance. It is reasonable to assume, that actual human reader performance is worse. In order to have a fair and realistic comparison between deep learning and human experts applying the Kellinghaus method a dedicated reading study needs to be conducted.

Thus, deep learning enables continuous and accurate predictions of chronological age. It provides an automated and thus scalable solution for radiological age assessment that might be improved with larger and optimized training datasets in the future.

In summary, this work includes four proof-of-concept studies that successfully addressed two clinical problems in radiology with machine learning. Hopefully, the now-proven ability of machine learning to differentiate colorectal polyps sparks further research to help advance therapy guidance in CT colonography cancer screenings. Also, the demonstrated continuous predictions of chronological age based on thoracic CT scans may inspire future radiological age assessment methods and improve the accuracy of age estimates.

8 | Bibliography

[1] M. Goldbaum, N. Katz, S. Chaudhuri, and M. Nelson, "Image Understanding for Automated Retinal Diagnosis", *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 756–760, Nov. 1989, ISSN: 0195-4210.

[2] L. Edenbrandt, B. Devine, and P. W. Macfarlane, "Neural networks for classification of ECG ST-T segments", *Journal of Electrocardiology*, vol. 25, no. 3, pp. 167–173, Jul. 1992, ISSN: 0022-0736. DOI: 10.1016/0022-0736(92)90001-G.

[3] M. Binder, A. Steiner, M. Schwarz, S. Knollmayer, K. Wolff, and H. Pehamberger, "Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: A pilot study", *British Journal of Dermatology*, vol. 130, no. 4, pp. 460–465, 1994, ISSN: 1365-2133. DOI: 10.1111/j.1365-2133.1994.tb03378.x.

[4] M. Beksaç, M. S. Beksaç, V. B. Tipi, H. A. Duru, M. Ü. Karakaş, and A. N. Çakar, "An artificial intelligent diagnostic system on differential recognition of hematopoietic cells from microscopic images", *Cytometry*, vol. 30, no. 3, pp. 145–150, 1997, ISSN: 1097-0320. DOI: 10.1002/(SICI)1097-0320(19970615)30:3<145::AID-CYTO5>3.0.CO;2-K.

[5] C. J. Haug and J. M. Drazen, "Artificial Intelligence and Machine Learning in Clinical Medicine, 2023", *New England Journal of Medicine*, vol. 388, no. 13, pp. 1201–1208, Mar. 2023, ISSN: 0028-4793. DOI: 10.1056/NEJMra2302038.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, ISSN: 1476-4687. DOI: 10.1038/323533a0.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12, Red Hook, NY, USA: Curran Associates Inc., Dec. 2012, pp. 1097–1105.

[8] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015, ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-015-0816-y.

[9] U.S. Food & Drug Administration. "Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices". (Oct. 2022), [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (visited on 06/21/2023).

[10] P. M. Cheng *et al.*, "Deep learning: An update for radiologists", *Radiographics : a review publication of the Radiological Society of North America, Inc*, vol. 41, no. 5, pp. 1427–1445, Sep. 2021, ISSN: 15271323. DOI: 10.1148/rg.2021200210.

[11] H. J. W. L. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach", *Nature Communications*, vol. 5, no. 1, p. 4006, Jun. 2014, ISSN: 2041-1723. DOI: 10.1038/ncomms5006.

[12] S. Grosu *et al.*, "Machine Learning-based Differentiation of Benign and Premalignant Colorectal Polyps Detected with CT Colonography in an Asymptomatic Screening Population: A Proof-of-Concept Study", *Radiology*, p. 202 363, Feb. 2021, ISSN: 0033-8419. DOI: 10.1148/radiol.2021202363.

[13] P. Wesp *et al.*, "Deep learning in CT colonography: Differentiating premalignant from benign colorectal polyps", *European Radiology*, vol. 32, no. 7, pp. 4749–4759, Jul. 2022, ISSN: 1432-1084. DOI: 10.1007/s00330-021-08532-2.

[14] P. Wesp *et al.*, "Automated localization of the medial clavicular epiphyseal cartilages using an object detection network: A step towards deep learning-based forensic age assessment", *International Journal of Legal Medicine*, Feb. 2023. DOI: 10.1007/s00414-023-02958-7.

[15] P. Wesp *et al.*, "Radiological age assessment based on clavicle ossification in ct: Enhanced accuracy through deep learning", *International Journal of Legal Medicine*, Jan. 2024. DOI: 10.1007/s00414-024-03167-6.

[16] R. L. Siegel *et al.*, "Colorectal cancer statistics, 2020", *CA: A Cancer Journal for Clinicians*, vol. 70, pp. 145–164, 3 May 2020, ISSN: 0007-9235. DOI: 10.3322/caac.21601.

[17] V. Kumar, A. K. Abbas, J. C. Aster, and S. L. Robbins, *Robbins Basic Pathology*. Elsevier, 2013.

[18] A. G. Zauber *et al.*, "Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths", *N Engl J Med*, vol. 366, pp. 687–696, 8 Feb. 2012. DOI: 10.1056/NEJMoa1100370.

[19] A. Graser *et al.*, "Comparison of CT colonography, colonoscopy, sigmoidoscopy and faecal occult blood tests for the detection of advanced adenoma in an average risk population", *Gut*, vol. 58, no. 2, pp. 241–248, Feb. 2009, ISSN: 1468-3288. DOI: 10.1136/gut.2008.156448.

[20] A. Laghi, E. Neri, and D. Regge, "Editorial on the european society of gastrointestinal endoscopy (esge) and european society of gastrointestinal and abdominal radiology (esgar) guideline on clinical indications for ct colonogra-

phy in the colorectal cancer diagnosis", *Radiologia Medica*, vol. 120, pp. 1021–1023, 11 Nov. 2015, ISSN: 18266983. DOI: 10.1007/s11547-015-0537-x.

[21] D. K. Rex *et al.*, "Colorectal cancer screening: Recommendations for physicians and patients from the u.s. multi-society task force on colorectal cancer", *Gastroenterology*, vol. 153, pp. 307–323, 1 Jul. 2017, ISSN: 15280012. DOI: 10.1053/j.gastro.2017.05.013.

[22] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 45, pp. 5–32, 2001, ISSN: 08856125. DOI: 10.1023/A:1010933404324.

[23] P. Lambin *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis", *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012, ISSN: 0959-8049, 1879-0852. DOI: 10.1016/j.ejca.2011.11.036.

[24] G. Chartrand *et al.*, "Deep learning: A primer for radiologists", *Radiographics : a review publication of the Radiological Society of North America, Inc*, vol. 37, no. 7, pp. 2113–2131, 2017, ISSN: 15271323. DOI: 10.1148/rg.2017170077.

[25] M. A. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

[26] J. J. V. Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype", *Cancer Research*, vol. 77, no. 21, e104–e107, Nov. 2017, ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-17-0339.

[27] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks", Mar. 2018. DOI: 10.1109/WACV.2018.00097.

[28] V. Kumar *et al.*, "Radiomics: The process and the challenges", *Magnetic Resonance Imaging*, Quantitative Imaging in Cancer, vol. 30, no. 9, pp. 1234–1248, Nov. 2012, ISSN: 0730-725X. DOI: 10.1016/j.mri.2012.06.010.

[29] A. H. Dachman *et al.*, "Effect of computer-aided detection for CT colonography in a multireader, multicase trial", *Radiology*, vol. 256, no. 3, pp. 827–35, Sep. 2010, ISSN: 0033-8419. DOI: 10.1148/radiol.10091890.

[30] S. Halligan *et al.*, "Incremental benefit of computer-aided detection when used as a second and concurrent reader of CT colonographic data: Multiobserver study", *Radiology*, vol. 258, no. 2, pp. 469–76, Feb. 2011, ISSN: 0033-8419. DOI: 10.1148/radiol.10100354.

[31] A. Vaswani *et al.*, *Attention Is All You Need*, Dec. 2017. DOI: 10.48550/arXiv.1706.03762.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-Resolution Image Synthesis with Latent Diffusion Models*, Apr. 2022. DOI: 10.48550/arXiv.2112.10752.

[33] United Nations. "The convention on the rights of the child". (1989), [Online]. Available: https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child (visited on 08/18/2023).

[34] The European Parliament Union and The Council Of The European. "Directive 2013/33/eu of the european parliament and of the council of 26 june

2013 laying down standards for the reception of applicants for international protection (recast)". (2013), [Online]. Available: http://data.europa.eu/eli/dir/2013/33/oj (visited on 08/18/2023).

[35] European Asylum Support Office, *EASO Practical Guide on Age Assessment*. Publications Office, 2018. DOI: 10.2847/292263.

[36] A. Schmeling, R. Dettmeyer, E. Rudolf, V. Vieth, and G. Geserick, "Forensic age estimation: Methods, certainty, and the law", *Deutsches Arzteblatt International*, vol. 113, no. 4, pp. 44–50, 2016, ISSN: 18660452. DOI: 10.3238/arztebl.2016.0044.

[37] M. Kellinghaus, R. Schulz, V. Vieth, S. Schmidt, and A. Schmeling, "Forensic age estimation in living subjects based on the ossification status of the medial clavicular epiphysis as revealed by thin-slice multidetector computed tomography", *International Journal of Legal Medicine*, vol. 124, no. 2, pp. 149–154, 2010, ISSN: 09379827. DOI: 10.1007/s00414-009-0398-8.

[38] M. Kellinghaus, R. Schulz, V. Vieth, S. Schmidt, H. Pfeiffer, and A. Schmeling, "Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans", *International Journal of Legal Medicine*, vol. 124, no. 4, pp. 321–325, 2010, ISSN: 09379827. DOI: 10.1007/s00414-010-0448-2.

[39] D. Wittschieber *et al.*, "The value of sub-stages and thin slices for the assessment of the medial clavicular epiphysis: A prospective multi-center CT study", *Forensic Science, Medicine, and Pathology*, vol. 10, no. 2, pp. 163–169, Jun. 2014, ISSN: 1556-2891. DOI: 10.1007/s12024-013-9511-x.

[40] M. Diligenti, S. Roychowdhury, and M. Gori, "Integrating prior knowledge into deep learning", in *16th IEEE International Conference on Machine Learning and Applications, ICMLA*, vol. 16, 2017, pp. 920–923, ISBN: 978-1-5386-1417-4. DOI: 10.1109/ICMLA.2017.00-37.

[41] M. J. Willemink *et al.*, "Preparing Medical Imaging Data for Machine Learning", *Radiology*, vol. 295, no. 1, pp. 4–15, 2020, ISSN: 15271315.

[42] W. C. Röntgen, "Ueber eine neue Art von Strahlen", *Annalen der Physik*, vol. 300, no. 1, pp. 1–11, 1898, ISSN: 15213889. DOI: 10.1002/andp.18983000102.

[43] D. Röntgen-Museum, "89008: Versuch: Durchleuchtung einer menschlichen Hand (22.12.1895)", Nov. 2021.

[44] A. M. Cormack, "Representation of a function by its line integrals, with some radiological applications", *Journal of Applied Physics*, vol. 34, no. 9, pp. 2822–2727, 1963, ISSN: 0021-8979. DOI: 10.1063/1.1729798.

[45] T. M. Buzug, *Computed Tomography*. Springer Berlin Heidelberg, 2008, ISBN: 978-3-540-39407-5. DOI: 10.1007/978-3-540-39408-2.

[46] J.-H. Grunert, *Strahlenschutz für Röntgendiagnostik und Computertomografie*. Springer Berlin Heidelberg, 2019, ISBN: 978-3-662-59274-8. DOI: 10.1007/978-3-662-59275-5.

[47] G. Poludniowski, A. Omar, R. Bujila, and P. Andreo, "Technical note: Spekpy v2.0—a software toolkit for modeling x-ray tube spectra", *Medical Physics*, vol. 48, pp. 3630–3637, 7 Jul. 2021, ISSN: 24734209. DOI: 10.1002/mp.14945.

[48] M. J. Berger *et al.*, *XCOM: Photon Cross Sections Database 8*, Version 2.0. National Institute of Standards and Technology, Oct. 2010.

[49] A. Einstein, "Über einen die erzeugung und verwandlung des lichtes betreffenden heuristischen gesichtspunkt", *Annalen der Physik*, vol. 322, pp. 132–148, 6 1905, ISSN: 15213889. DOI: 10.1002/andp.19053220607.

[50] T. Laubenberger and J. Laubenberger, *Technik der medizinischen Radiologie: Diagnostik, Strahlentherapie, Strahlenschutz ; für Ärzte, Medizinstudenten und MTRA ; [mit 71 Tabellen]*. Deutscher Ärzteverlag, 1999, ISBN: 978-3-7691-1132-3.

[51] K. Rajendran *et al.*, "First clinical photon-counting detector ct system: Technical evaluation", *Radiology*, vol. 303, pp. 130–138, 1 Apr. 2022, ISSN: 15271315. DOI: 10.1148/RADIOL.212579.

[52] L. A. Shepp and B. F. Logan, "The fourier reconstruction of a head section", *IEEE Transactions on Nuclear Science*, vol. 21, pp. 21–43, 3 Jun. 1974. DOI: 10.1109/TNS.1974.6499235.

[53] E. Krestel, *Imaging Systems in Medical Diagnostics*, Second Edition. Germany: Siemens, 1990.

[54] M. Hofer, *CT-Kursbuch*. Düsseldorf: Didimed, 2000.

[55] H. Brenner, C. Stock, and M. Hoffmeister, "Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: Systematic review and meta-analysis of randomised controlled trials and observational studies", *BMJ*, vol. 348, Apr. 2014, ISSN: 17561833. DOI: 10.1136/bmj.g2467.

[56] M. P. van der Meulen *et al.*, "Colorectal cancer: Cost-effectiveness of colonoscopy versus ct colonography screening with participation rates and costs", *Radiology*, vol. 287, pp. 901–911, 3 2018.

[57] B. D. Pooler, D. H. Kim, J. M. Weiss, K. A. Matkowskyj, and P. J. Pickhardt, "Colorectal polyps missed with optical colonoscopy despite previous detection and localization with ct colonography", *Radiology*, vol. 278, pp. 422–429, 2 2016.

[58] E. M. Stoop *et al.*, "Participation and yield of colonoscopy versus non-cathartic CT colonography in population-based screening for colorectal cancer: A randomised controlled trial", *The Lancet Oncology*, vol. 13, no. 1, pp. 55–64, Jan. 2012, ISSN: 14702045. DOI: 10.1016/S1470-2045(11)70283-2.

[59] L. Sali *et al.*, "Reduced and Full-Preparation CT Colonography, Fecal Immunochemical Test, and Colonoscopy for Population Screening of Colorectal Cancer: A Randomized Trial", *Journal of the National Cancer Institute*, vol. 108, no. 2, djv319, Feb. 2016, ISSN: 0027-8874, 1460-2105. DOI: 10.1093/jnci/djv319.

[60] A. Schmeling *et al.*, *Aktualisierte Empfehlungen für Altersschätzungen bei Lebenden im Strafverfahren*.

[61] W. W. Greulich and S. I. Pyle, *Radiographic Atlas of Skeletal Development of the Hand and Wrist*. Stanford, CA, USA: Stanford University Press, 1959.

[62] A. Demirjian, H. Goldstein, and J. M. Tanner, "A New System of Dental Age Assessment", *Human Biology*, vol. 45, no. 2, pp. 211–227, 1973, ISSN: 0018-7143.

[63] A. Schmeling, R. Schulz, W. Reisinger, M. Mühler, K. D. Wernecke, and G. Geserick, "Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography", *International Journal of Legal Medicine*, vol. 118, no. 1, pp. 5–8, 2004, ISSN: 09379827. DOI: 10.1007/s00414-003-0404-5.

[64] P. S. Dahlberg *et al.*, "A systematic review of the agreement between chronological age and skeletal age based on the Greulich and Pyle atlas", *European Radiology*, vol. 29, no. 6, pp. 2936–2948, Jun. 2019, ISSN: 1432-1084. DOI: 10.1007/s00330-018-5718-2.

[65] A. Schmeling, W. Reisinger, D. Loreck, K. Vendura, W. Markus, and G. Geserick, "Effects of ethnicity on skeletal maturation: Consequences for forensic age estimations", *International Journal of Legal Medicine*, vol. 113, no. 5, pp. 253–258, 2000, ISSN: 09379827. DOI: 10.1007/s004149900102.

[66] J. M. Lewis and D. R. Senn, "Dental age estimation utilizing third molar development: A review of principles, methods, and population studies used in the United States", *Forensic Science International*, vol. 201, no. 1-3, pp. 79–83, Sep. 2010, ISSN: 03790738. DOI: 10.1016/j.forsciint.2010.04.042.

[67] H. H. Thodberg, "An Automated Method for Determination of Bone Age", *The Journal of Clinical Endocrinology & Metabolism*, vol. 94, no. 7, pp. 2239–2244, Jul. 2009, ISSN: 0021-972X, 1945-7197. DOI: 10.1210/jc.2008-2474.

[68] S. H. Tajmir *et al.*, "Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability", *Skeletal Radiology*, vol. 48, no. 2, pp. 275–283, Feb. 2019, ISSN: 1432-2161. DOI: 10.1007/s00256-018-3033-2.

[69] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane, "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming", *Artificial Intelligence in Design '96*, J. S. Gero and F. Sudweeks, Eds., pp. 151–170, 1996. DOI: 10.1007/978-94-009-0279-4_9.

[70] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers", *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, Jul. 1959, ISSN: 0018-8646. DOI: 10.1147/rd.33.0210.

[71] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2009, ISBN: 978-0-387-31073-2.

[72] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R* (Springer Texts in Statistics). New York, NY: Springer US, 2021, ISBN: 978-1-07-161417-4.

[73] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[74] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, ISSN: 1476-4687. DOI: 10.1038/nature14539.

[75] L. Breiman, "Bagging predictors", *Machine learning*, vol. 24, pp. 123–140, 1996, ISSN: 0885-6125. DOI: 10.1007/BF00058655.

[76] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees", *Neural computation*, vol. 9, no. 7, pp. 1545–1588, 1997. DOI: 10.1162/neco.1997.9.7.1545.

[77] T. K. Ho, "The random subspace method for constructing decision forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832–844, 1998. DOI: 10.1109/34.709601.

[78] L. Breiman, *Out-of-bag Estimation*, 1996.

[79] G. Louppe, *Understanding Random Forests: From Theory to Practice*, Jun. 2015. DOI: 10.48550/arXiv.1407.7502.

[80] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer, "High-Dimensional Variable Selection for Survival Data", *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 205–217, Mar. 2010, ISSN: 0162-1459. DOI: 10.1198/jasa.2009.tm08622.

[81] J. Song, Y. Yin, H. Wang, Z. Chang, Z. Liu, and L. Cui, "A review of original articles published in the emerging field of radiomics", *European Journal of Radiology*, vol. 127, p. 108 991, Jun. 2020, ISSN: 0720-048X. DOI: 10.1016/j.ejrad.2020.108991.

[82] A. Zwanenburg and M. Vallières, "The image biomarker standardisation initiative", *arXiv preprint*, pp. 1–168, 2019. DOI: 10.17195/candat.2016.08.1.

[83] G. Cybenko, "Approximation by superpositions of a sigmoidal function", *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, Dec. 1989, ISSN: 1435-568X. DOI: 10.1007/BF02551274.

[84] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943, ISSN: 1522-9602. DOI: 10.1007/BF02478259.

[85] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization Methods for Large-Scale Machine Learning*, Feb. 2018. DOI: 10.48550/arXiv.1606.04838.

[86] H. Robbins and S. Monro, "A Stochastic Approximation Method", *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, Sep. 1951, ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177729586.

[87] J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function", *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, 1952, ISSN: 0003-4851.

[88] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization", in *3rd International Conference on Learning Representations (ICLR)*, vol. 3, San Diego, CA, USA, 2015.

[89] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, Apr. 1980, ISSN: 1432-0770. DOI: 10.1007/BF00344251.

[90] Y. LeCun *et al.*, "Backpropagation Applied to Handwritten Zip Code Recognition", *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989, ISSN: 0899-7667. DOI: 10.1162/neco.1989.1.4.541.

[91] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, ISSN: 1558-2256. DOI: 10.1109/5.726791.

[92] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014, ISSN: 10636919. DOI: 10.1109/CVPR.2014.81.

[93] R. Girshick, "Fast R-CNN", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1440–1448.

[94] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", ser. European Conference on Computer Vision (ECCV), 2014. DOI: 10.1109/TPAMI.2016.2577031.

[95] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You Only Look Once: Unified, Real-Time Object Detection*, May 2016. DOI: 10.48550/arXiv.1506.02640.

[96] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020, ISSN: 19393539. DOI: 10.1109/TPAMI.2018.2858826.

[97] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection", ser. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, 2017, pp. 2117–2125.

[98] M. Everingham and J. Winn. "The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Development Kit". (2010), [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/voc2010/devkit_doc_08-May-2010.pdf (visited on 08/18/2023).

[99] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge", *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010, ISSN: 09205691. DOI: 10.1007/s11263-009-0275-4.

[100] B. Sahiner *et al.*, "Deep learning in medical imaging and radiation therapy", *Medical Physics*, vol. 46, no. 1, e1–e36, 2019. DOI: 10.1002/mp.13264.

[101] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A

review", *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, Jan. 2021, ISSN: 0269-2821, 1573-7462. DOI: 10.1007/s10462-020-09854-1.

[102] D. Shen, G. Wu, and H.-I. Suk, "Deep Learning in Medical Image Analysis", *Annu Rev Biomed Eng*, no. 19, pp. 221–248, 2017. DOI: 10.1146/annurev-bioeng-071516-044442.

[103] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine", *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019, ISSN: 0028-4793. DOI: 10.1056/nejmra1814259.

[104] C. D. Becker, E. Kotter, L. Fournier, and L. Martí-Bonmatí, "Current practical experience with artificial intelligence in clinical radiology: A survey of the European Society of Radiology", *Insights into Imaging*, vol. 13, no. 1, Dec. 2022, ISSN: 18694101. DOI: 10.1186/s13244-022-01247-y.

[105] A. Reinke *et al.*, *Common Limitations of Image Processing Metrics: A Picture Story*, Jul. 2022. DOI: 10.48550/arXiv.2104.05642.

[106] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence", *Nature*, vol. 616, no. 7956, pp. 259–265, Apr. 2023, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-05881-4.

[107] R. Bommasani *et al.*, *On the Opportunities and Risks of Foundation Models*, Jul. 2022. DOI: 10.48550/arXiv.2108.07258.

[108] E. Rudolf, J. Kramer, S. Schmidt, V. Vieth, I. Winkler, and A. Schmeling, "Anatomic shape variants of extremitas sternalis claviculae as collected from sternoclavicular thin-slice CT-studies of 2820 male borderline-adults", *International Journal of Legal Medicine*, vol. 133, no. 5, pp. 1517–1528, Sep. 2019, ISSN: 1437-1596. DOI: 10.1007/s00414-019-02065-6.

[109] J. De Tobel *et al.*, "Staging Clavicular Development on MRI: Pitfalls and Suggestions for Age Estimation", *Journal of Magnetic Resonance Imaging*, vol. 51, no. 2, pp. 377–388, 2020, ISSN: 1522-2586. DOI: 10.1002/jmri.26889.

A | Appendix

A.1 Complementing Publication I

International Journal of Legal Medicine
https://doi.org/10.1007/s00414-024-03167-6

ORIGINAL ARTICLE

Radiological age assessment based on clavicle ossification in CT: enhanced accuracy through deep learning

Philipp Wesp[1,2] · Balthasar Maria Schachtner[1] · Katharina Jeblick[1,3] · Johanna Topalis[1] · Marvin Weber[4] · Florian Fischer[5] · Randolph Penning[5] · Jens Ricke[1] · Michael Ingrisch[1,2] · Bastian Oliver Sabel[1]

Received: 13 July 2023 / Accepted: 16 January 2024

Abstract

Background Radiological age assessment using reference studies is inherently limited in accuracy due to a finite number of assignable skeletal maturation stages. To overcome this limitation, we present a deep learning approach for continuous age assessment based on clavicle ossification in computed tomography (CT).

Methods Thoracic CT scans were retrospectively collected from the picture archiving and communication system. Individuals aged 15.0 to 30.0 years examined in routine clinical practice were included. All scans were automatically cropped around the medial clavicular epiphyseal cartilages. A deep learning model was trained to predict a person's chronological age based on these scans. Performance was evaluated using mean absolute error (MAE). Model performance was compared to an optimistic human reader performance estimate for an established reference study method.

Results The deep learning model was trained on 4,400 scans of 1,935 patients (training set: mean age = 24.2 years ± 4.0, 1132 female) and evaluated on 300 scans of 300 patients with a balanced age and sex distribution (test set: mean age = 22.5 years ± 4.4, 150 female). Model MAE was 1.65 years, and the highest absolute error was 6.40 years for females and 7.32 years for males. However, performance could be attributed to norm-variants or pathologic disorders. Human reader estimate MAE was 1.84 years and the highest absolute error was 3.40 years for females and 3.78 years for males.

Conclusions We present a deep learning approach for continuous age predictions using CT volumes highlighting the medial clavicular epiphyseal cartilage with performance comparable to the human reader estimate.

Keywords X-Ray computed tomography · Age determination by skeleton · Deep learning · Sternoclavicular joint · Forensic medicine

Background

Radiological age assessment is a method that examines certain physiological properties in radiographic or computed tomography (CT) images to estimate a person's chronological age [1, 2]. In this study, we explore a potential approach to enhance radiological age assessment based on clavicle bone ossification through deep learning.

Importance of age

In many countries, age governs the relationship between individuals and the state. Changes in age can lead to the acquisition of rights and obligations, such as emancipation, employment, criminal responsibility, sexual relation,

✉ Philipp Wesp
philipp.wesp@med.uni-muenchen.de

1 Department of Radiology, LMU University Hospital, LMU Munich, Marchioninistraße 15, 81377 Munich, Germany

2 Munich Center for Machine Learning (MCML), Geschwister-Scholl-Platz 1, 80539 Munich, Germany

3 Comprehensive Pneumology Center (CPC-M), Member of the German Center for Lung Research (DZL), Munich, Max-Lebsche-Platz 31, 81377 Munich, Germany

4 Institute of Informatics, LMU Munich, Oettingenstraße 67, 80538 Munich, Germany

5 Institute of Forensic Medicine, LMU Munich, Nußbaumstraße 26, 80336 Munich, Germany

Published online: 30 January 2024

consent for marriage, or military service [3]. Thus, age is a critical component of a person's identity, particularly for children. The United Nations Convention on the Rights of the Child (CRC, Article 1) [4] and the EU acquis (Directive 2013/33/EU, Article 2(d)) [5] define a child as any person below the age of 18. States and authorities have specific age-related obligations under the CRC that include: registration of the child after birth, respecting the right of the child to preserve his or her identity, and speedily re-establish his or her identity in the case that some or all elements of the child's identity have been deprived [3]. In cases where a person's age is unknown or in serious doubt, a state may need to assess the age, e.g., to determine whether they are an adult or a child. The European Union Agency for Asylum (EUAA) recommends using the least intrusive age assessment method possible, gradually implementing more invasive methods if necessary, and selecting the most accurate method while documenting the margin of error [3]. Radiological age assessment is one such method and its accuracy may be improved using deep learning. Other non-binding recommendations from local expert panels exist, e.g., from the Working Group for Forensic Age Diagnostics of the German Society for Forensic Medicine (AGFAD).[1]

Reference study-based radiological age assessment

Radiological age assessment is based on examinations of body parts that capture the skeletal development of the person whose age is unknown, such as the carpal bones, the molars, or the clavicles [1]. In this study, we focus on the ossification status of the medial clavicular epiphyseal cartilages, as they are the last maturing bone structures in the human body, and enable the estimation of a wide range of ages, from teenagers to young adolescents and adults [6]. Typically, atlas methods [2] or reference study methods [7–9] are applied for age assessment, where the age of the examined person is assumed to be similar to the reference person or case group with similar skeletal maturation.

However, these methods have several limitations. First, the number of case groups is finite, e.g. n = 9 in [7–9], which limits the accuracy of age estimates. Second, age differences between members of the same case group can be large, e.g., up to 14.2 years [7], leading to high uncertainties. Third, expanding control groups is challenging because the assessment of the ossification stage by experts is time-consuming. Finally, these methods are subject to intra- and inter-reader variability [10, 11].

[1] https://www.dgrm.de/forensische-altersdiagnostik/empfehlungen (2023/11/28).

Deep learning-based radiological age assessment

A promising tool for more accurate radiological age assessment via the clavicle bones is deep learning. It has been successfully applied in a variety of computer vision tasks in medical imaging [12] including radiological age assessment through dental radiographs [13], knee MRIs [14], and more [15]. The large amounts of data required to train a deep network for age assessment [16]—medical images including clavicles and sternum, along with the corresponding age information—are abundant in many hospitals and can be accessed retrospectively through their picture archiving and communication systems (PACS). Furthermore, data from institutions in different locations can be combined to form a dataset that is representative of the global population as well as possible. Finally, feed-forward deep learning models are deterministic as the same input image always results in the same output and age predictions do not suffer from intra- or inter-rater variability. This might be an advantage when considering which method should be deployed in potential legal scenarios.

Therefore, we (a) propose a deep learning approach to predict the chronological age based on CT image volumes of the medial clavicular epiphyseal cartilage and (b) compare it to a favorable human reader performance estimate for the reference study method of Kellinghaus et al. [7, 8]. It is widely acknowledged in conventional practice that the classification of stage 3b in males and stage 3c in females following the Kellinghaus method suggests a minimum age of 18 years or above.

Methods

Retrospective data collection

This retrospective study was approved by the institutional review board (Ethics Committee, Medical Faculty, LMU Munich) and the requirement for written informed consent was waived. CT scans were collected retrospectively from the PACS of LMU Munich's University Hospital. We specifically searched for chest CT scans of persons between the ages of 15.0 and 30.0 years, with documented sex, reimbursed by a recognized health-insurance provider (state-mandated or private), acquired during the clinical routine for all purposes between 2017 to 2020. To ensure truthful age information we excluded scans issued and paid for by state agencies, which among other things excludes requests for forensic age assessments. Age was calculated as the number of days between the date of birth and the date of examination. The selected age range covers a broad spectrum of skeletal developmental stages of the medial clavicular

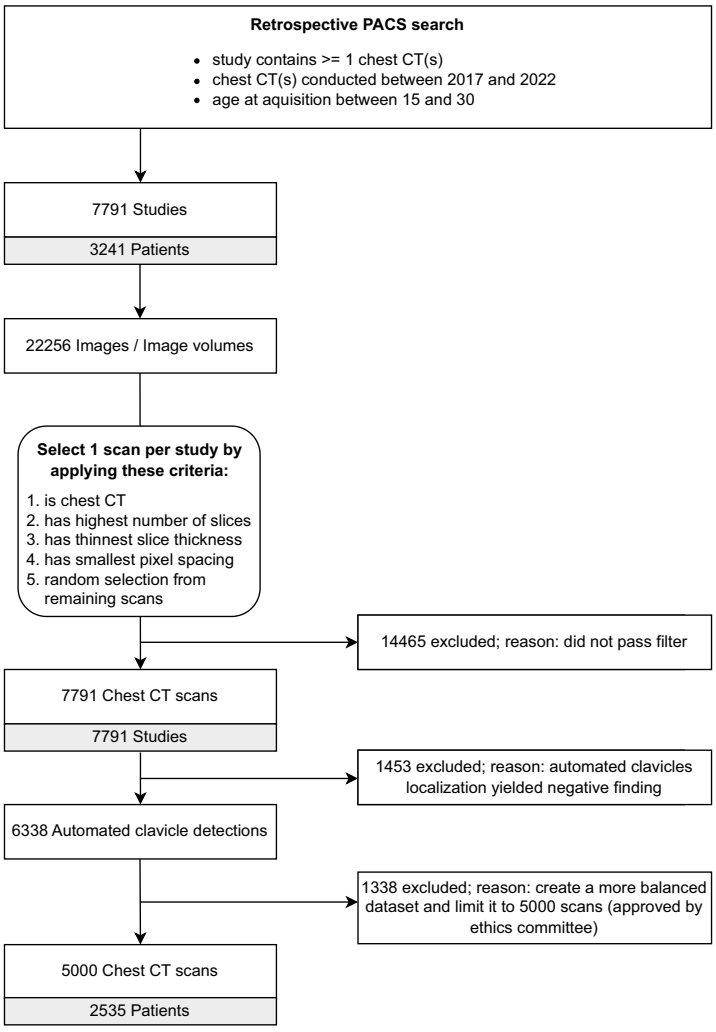

Fig. 1 CT scan inclusion diagram. Flow diagram of the selection process from study identification in the picture archiving and communication system (PACS) to the chest CT scans in the dataset

epiphyseal cartilages [17]. One scan per study was selected based on multiple criteria specified in the flow diagram in Fig. 1, which summarizes the entire data collection process.

Deep learning model

A schematic overview of the deep learning approach for radiological age assessment is shown in Fig. 2. We express age assessment as a regression analysis where the dependent variable (age) is a scalar, which is estimated based on a feature (CT scan), by a deep learning model. The model in this study was an ensemble [18] of 20 deep neural networks (deep ensemble) that share the same architecture and training process. The mean of the predictions from the 20 ensemble members was used as the ensemble prediction. The architecture was adapted from the popular ResNet-18 [19], where we replaced the two-dimensional convolutions with three-dimensional convolutions to enable processing CT volume inputs, and added a second input to process sex information.

Prior to model training, the collected CT scans were preprocessed (described in detail in the supplement) including an automated localization of the clavicles [20]. This localization also served as a filter for chest CT scans that do not include the clavicles or scans wrongly labelled as chest CT. Next, the dataset was split into a training, a validation, and a test set. Validation and test set were sampled to include not more than one CT scan of the same person and to have the same equal number of samples per age (bin size = 1 year) and sex. All remaining samples from persons not in the

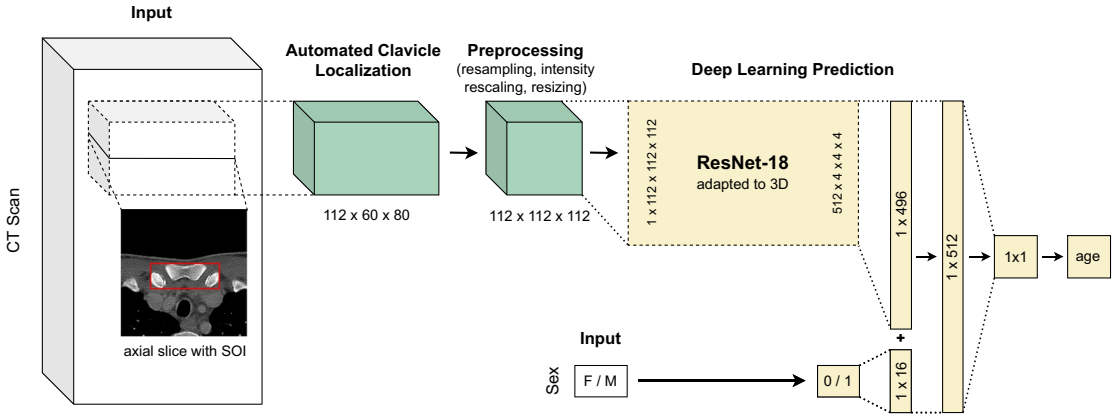

Fig. 2 Deep learning-based radiological age assessment. Schematic visualization of the proposed approach for deep learning-based radiological age assessment. First, the CT scan is cropped around the automatically localized structures of interest (SOIs), which are the medial clavicular epiphyseal cartilages. Second, the scan undergoes several preprocessing steps which include resampling, intensity rescaling, and resizing. Finally, the adapted three-dimensional ResNet-18 predicts chronological age based on the preprocessed scan. Additionally, sex information is incorporated into the approach by fusing it with the image embedding before the last fully connected layer. While the figure only depicts a single network, the deep learning approach uses a deep ensemble consisting of 20 uniquely trained networks

validation or test set were used as the training set. No person is part of more than one set. The deep ensemble was trained on the training set, and training progress was monitored using the validation set. Model performance was evaluated by measuring the absolute error of model predictions for the test set. Details regarding the dataset split, model, and training are provided in the supplement.

Abstention-performance trade-off

We applied the estimated predictive uncertainty of the deep ensemble to identify samples with a potentially high prediction error. The standard deviation (SD) of the predictions made by the ensemble members for a given input served as the respective uncertainty estimate [21]. In an abstention-performance trade-off, we abstain from predictions for the fraction of samples with the highest measured uncertainties (abstention rate) to improve average performance for the remaining samples. For example, in a trade-off with an abstention rate of 20%, we rank all predictions by predictive uncertainty and analyze only the top 80% of samples with the lowest uncertainty. This allows the machine learning model to say "I don't know" [22] in cases where it is unsure, instead of forcing an answer at all costs.

Optimistic human reader performance estimate

To classify the performance of our deep learning model, we calculated an optimistic human reader performance estimate for the radiological age assessment of Kellinghaus et al. [7, 8]. This method is based on 9 clavicle ossification stages, with three major stages (1, 4, and 5) and 6 substages (2a—2c and 3a—3c). They range from no ossification of the ossification center (stage 1) to complete fusion of the epiphyseal cartilage (stage 5). An individual's age is estimated by first determining the ossification stage in a radiological examination [7, 8]. Next, the age is derived from the age distribution of a case group of known age and with the same ossification stage and sex.

The human reader estimate assumes a best-case scenario in which (a) the descriptive ossification stage statistics described in [7, 8] are derived from a cohort that is representative of all individuals, in particular, our test set, (b) age in each stage follows a normal distribution and (c) trained reviewers always assess the correct ossification stage. Under these conditions the HRE provides the lower limit for the absolute error that can be achieved with the reference study method when applied to a person with a certain true age x (Fig. 3).

For a given age x we first calculated the absolute difference to the mean age M of each ossification stage s:

$$|x - M(s)|$$

For example, for a 21.00 year old male, these differences are 7.72 years for stage 1 (M = 13.28 years), 3.60 years for stage 2a (M = 17.40 years), 2.80 years for stage 2b (M = 18.20 years), 2.40 years for stage 2c (M = 18.60 years), 2.00 years for stage 3a (M = 19.00 years), 0.10 years for stage 3b (M = 21.10 years), 1.90 years

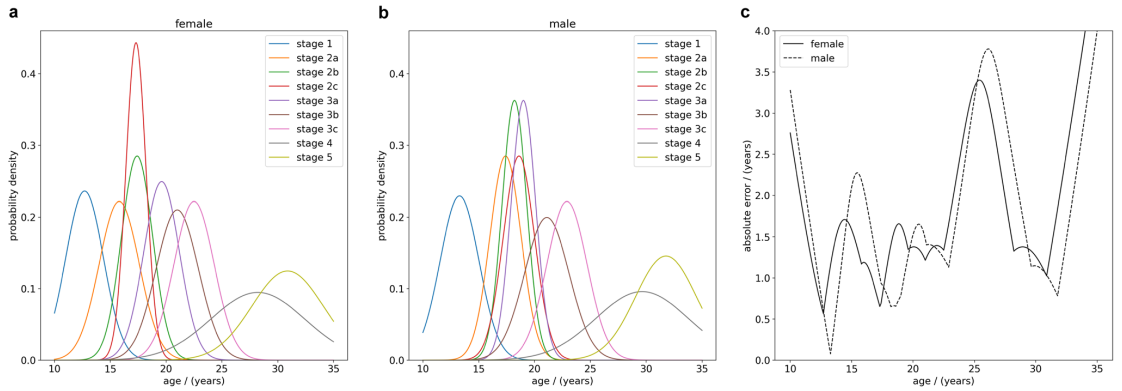

Fig. 3 Optimistic human reader performance estimate. The left and center panels display the probability density of a person being in a certain ossification stage, based on normal distributions described in [7, 8], for (**a**) females and (**b**) males between the ages of 10 and 35 years. The right panel (**c**) shows the best-case mean absolute error estimate of predicted ages for true ages between 10 and 35 years when applying the radiological reference study method for age assessment of Kellinghaus et al. [7, 8]

for stage 3c (M = 22.90 years), 8.63 years for stage 4 (M = 29.63 years), 10.77 years for stage 5 (M = 31.77 years).

Next, we calculated the probability density $p_s(x)$ (Fig. 3) for a person with the true chronological age x to be in ossification stage s based on normal distributions calculated from the provided mean and SD values. The probabilities were normalized such that

$$\sum_{s=1} p_s(x) = 1.$$

It is important to note, that two persons of the same chronological age can be in two different ossification stages. In the example of a 21.00 year old male, these probabilities are $p_1 = 2.45 \times 10^{-5}$, $p_{2a} = 2.10 \times 10^{-2}$, $p_{2b} = 2.86 \times 10^{-2}$, $p_{2c} = 1.32 \times 10^{-1}$, $p_{3a} = 1.40 \times 10^{-1}$, $p_{3b} = 4.01 \times 10^{-1}$, $p_{3c} = 2.55 \times 10^{-1}$, $p_4 = 2.24 \times 10^{-2}$, and $p_5 = 1.29 \times 10^{-4}$.

The probability densities $p_s(x)$ were multiplied by the absolute difference to the mean age.

$$p_s(x) \cdot |x - M(s)|$$

The sum of these products for all ossification stages yielded the absolute error of the reference study method for a person with the true age x:

$$AE(x) = \sum_{s=1} |M(s) - x| p_s(x)$$

In the example of the 21.00 year old male, the AE is 1.64 years. The MAE of the reference study method for all individuals in the test set was then given by:

$$MAE = \frac{1}{|X_{Test}|} \sum_{x \in X_{Test}} AE(x).$$

Classical expert reader age assessment

A senior radiologist and expert in the field conducted a manual reading of a small subset of the test set, comprising 50 randomly sampled test set scans. The reading followed the Kellinghaus method [7, 8] and assessed the ossification stages 1, 2a, 2b, 2c, 3a, 3b, 3c, 4, and 5. The mean age value of each stage of the respective sex was used as age prediction for the manual reading.

Results

Dataset

A retrospective search in our hospital's PACS identified 7,791 studies conducted between 2017 and 2020 on 3,241 patients that involved at least one chest CT scan with a recorded age at acquisition between 15 and 30 years, documented sex, and recognized health insurance provider (state-mandated or private). The 7,791 studies included 22,256 images or image volumes. Some studies included more than one chest CT scan that would have been suitable for analysis. After scan selection (Fig. 1), the final dataset consisted of 5,000 chest CT scans from 2,535 patients (mean age = 24.2 ± 4.0 years), with 44% (1,103/2,535) females. The training set consisted of

Table 1 Documentation of the number of patients and CT scans in the total dataset, as well as in the training, validation, and test set

Set	**Total**		**Training**	
Patients	f	m	f	m
	1103 (44%)	1432 (56%)	803 (41%)	1132 (59%)
	2535		1935	
CT scans	5000		4400	
Set	**Validation**		**Test**	
Patients	f	m	f	m
	150 (50%)	150 (50%)	150 (50%)	150 (50%)
	300		300	
CT scans	300		300	

4,400 scans from 1,935 patients, with 41% (803/1,935) female. The validation and test set were independent and both included 300 scans from 300 patients (both: mean age = 22.5 ± 4.4 years), 10 scans per age (bin size = 1 year), and sex, with 50% (150/300) being female. All datasets are summarized in Table 1 and their age distribution is shown in supplementary Figure S3.

Deep learning-based radiological age assessment

The deep ensemble model (Fig. 2) was trained using the training data and training was monitored using the validation data. The model's performance was evaluated on the test data. The results showed a mean absolute error (MAE) of 1.65 years (standard deviation (SD) = 0.53) for all patients, 1.69 years (SD = 0.53) for female patients, and 1.62 years (SD = 0.54) for male patients. The best prediction for a female individual had an absolute error of 0.003 years (true age = 18.604 years), while the best prediction for a male had an absolute error of 0.005 years (true age = 25.142 years). The corresponding input CT scans are displayed in Fig. 4 and show no medical abnormalities. The worst prediction for a female had an absolute error of 6.40 years (true age = 15.29 years). The corresponding CT showed a fish mouth shape variant with concavely configured clavicle ends in the left clavicle (Fig. 4). Shape variants near the sternal ends of the clavicle occur frequently and severely limit assessability [23, 24]. The worst prediction for a male had an absolute error of 7.32 years (true age = 19.20 years). A CT examination revealed that the individual had osteolysis

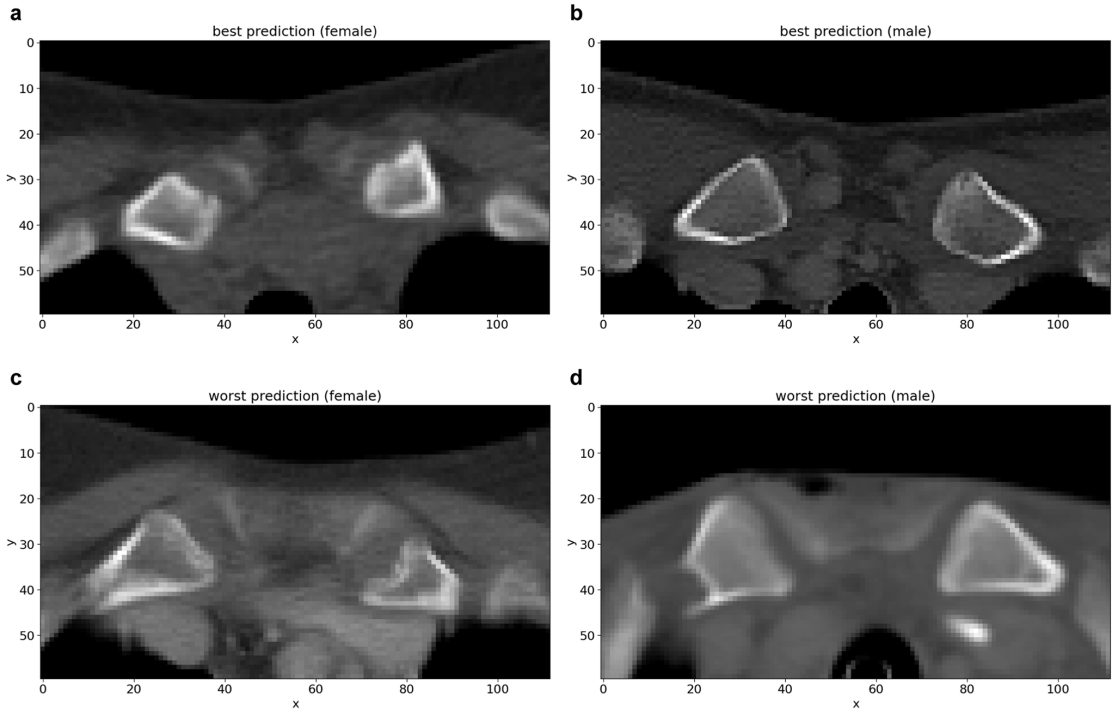

Fig. 4 Test set input examples. Selected axial slices of the preprocessed CT scans of (**a**) the best female, (**b**) best male, (**c**) worst female, and (**d**) worst male deep learning prediction for age. The worst predictions show (**c**) a "fish mouth configuration of the left clavicle" and a (**d**) osteolytic lesion of the right clavicle

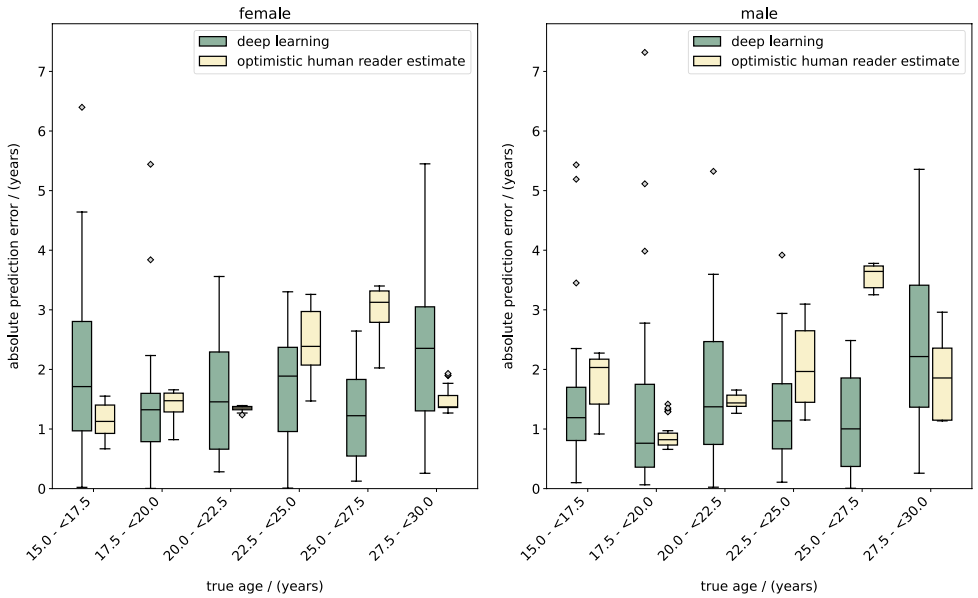

Fig. 5 Radiological age assessment results. Absolute prediction error of the (green) deep learning approach and the (yellow) optimistic human reader performance estimate for radiological age assessment of (left panel) females and (right panel) males between 15 and 30 years. The boxes extend from the lower to the upper quartile age values of each bin, with a line at the median. The whiskers extend from the boxes to 1.5 × interquartile range (IQR) (Q3—Q1) in each direction. Flier points are age values past the end of the whiskers

Table 2 Radiological age assessment results for female persons in the test set using deep learning (DL) and optimistic human reader performance estimate (HRE). The table displays the mean absolute error (MAE), maximum absolute error (max error), and 90th percentile absolute error (p90 error) in years. The number of individuals in each age group was n = 10

Results for female subjects						
Age	Deep learning			Human reader estimate		
	MAE	Max err	p90 err	MAE	Max err	p90 err
15.0- < 16.0	2.85	6.40	4.82	1.41	1.55	1.51
16.0- < 17.0	1.68	2.90	2.78	1.02	1.15	1.14
17.0- < 18.0	1.16	3.84	1.80	0.90	1.23	1.13
18.0- < 19.0	0.94	1.88	1.76	1.57	1.66	1.66
19.0- < 20.0	1.70	5.44	2.55	1.45	1.55	1.55
20.0- < 21.0	1.61	2.48	2.44	1.33	1.38	1.38
21.0- < 22.0	1.18	3.56	2.42	1.34	1.39	1.39
22.0- < 23.0	2.01	3.52	3.14	1.49	1.73	1.70
23.0- < 24.0	1.52	3.30	2.92	2.24	2.56	2.52
24.0- < 25.0	1.72	2.61	2.43	3.02	3.26	3.25
25.0- < 26.0	1.24	2.64	2.36	3.34	3.40	3.39
26.0- < 27.0	1.15	2.03	1.94	2.88	3.20	3.06
27.0- < 28.0	1.45	3.05	2.98	1.85	2.35	2.06
28.0- < 29.0	2.55	5.45	4.21	1.37	1.39	1.38
29.0- < 30.0	2.53	4.24	3.78	1.34	1.37	1.37

in the right clavicle, presumably as a manifestation of an underlying malignant disease (Fig. 4). The distribution of absolute errors by age (bin width = 2.5 years) is shown separately for male and female patients in Fig. 5. The MAE, maximum absolute error (max error), and the 90th percentile absolute error (p90 error) for each age (bin width = 1 year) are reported in Table 2 for female individuals and Table 3 for male individuals.

Table 3 Radiological age assessment results for male persons in the test set using deep learning (DL) and optimistic human reader performance estimate (HRE). The table displays the mean absolute error (MAE), maximum absolute error (max error), and 90th percentile absolute error (p90 error) in years. The number of individuals in each age group was n = 10

Results for male subjects						
Age	Deep learning			Human reader estimate		
	MAE	Max err	p90 err	MAE	Max err	p90 err
15.0- < 16.0	1.29	2.35	1.85	2.18	2.27	2.27
16.0- < 17.0	1.74	5.43	3.65	1.7	2.06	2.04
17.0- < 18.0	1.1	5.19	2.08	0.9	1.21	0.99
18.0- < 19.0	1.26	5.11	2.53	0.7	0.77	0.75
19.0- < 20.0	2.18	7.32	4.32	1.12	1.42	1.36
20.0- < 21.0	1.44	3.04	2.42	1.6	1.65	1.65
21.0- < 22.0	1.86	3.35	3.34	1.42	1.47	1.45
22.0- < 23.0	1.72	5.32	3.77	1.25	1.32	1.31
23.0- < 24.0	1.34	3.92	2.76	1.66	2.14	1.99
24.0- < 25.0	1.48	2.94	2.82	2.69	3.1	3.01
25.0- < 26.0	0.96	2.26	1.91	3.54	3.76	3.7
26.0- < 27.0	1.11	2.43	1.91	3.68	3.78	3.78
27.0- < 28.0	1.57	3.37	2.57	2.97	3.33	3.32
28.0- < 29.0	2.57	4.75	3.97	2.01	2.39	2.28
29.0- < 30.0	2.67	5.36	4.48	1.19	1.51	1.24

Optimistic human reader radiological age assessment

The human reader estimate for the radiological age assessment method of Kellinghaus et al. [7, 8] was applied to the test set. The results showed a MAE of 1.84 years (SD = 0.84 years) overall, 1.77 years (SD = 0.74 years) for female individuals, and 1.91 years (SD = 0.92 years) for male individuals. The distribution of absolute errors by age (bin width = 2.5 years) is shown separately for male and female individuals in Fig. 5. The MAE, max error, and p90 error for each age (bin width = 1 year) are reported in Table 2 for female patients and Table 3 for male patients.

Classical expert reader age assessment

The manual age assessment of 50 randomly sampled test set scans by an expert in the field following the method of Kellinghaus et al. [7, 8] yielded a MAE of 1.97 years (SD = 1.48 years). For comparison, the deep learning model achieved a MAE of 1.44 years (SD = 0.95 years) on the same subset.

Abstention-performance trade-off

In a separate analysis, we applied an abstention-performance trade-off to the deep learning model predictions, i.e. we did not take results from samples with the highest predictive uncertainties into account. MAE, max error, and p90 error for abstention rates ranging from 0% (all samples evaluated, no abstention) to 100% (no samples evaluated) are shown in

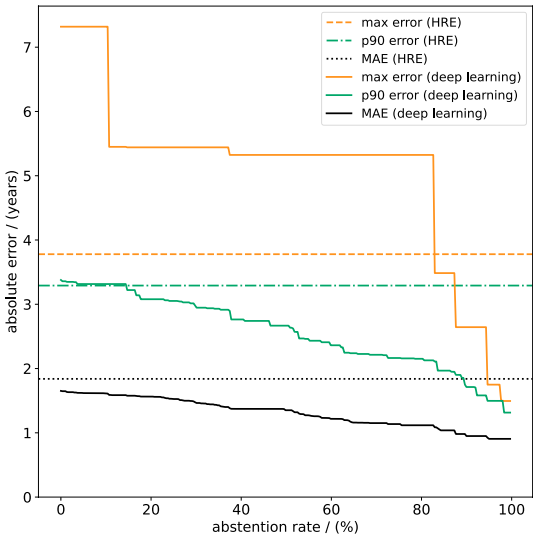

Fig. 6 Abstention-performance trade-off. The abstention-performance trade-off for deep learning-based radiological age assessment, where we abstain from analysis for predictions with the highest predictive uncertainties. Optimistic human reader performance estimate (HRE) results are included for reference. Increasing abstention rates lead to an improved deep learning mean absolute error (MAE), maximum absolute error (max error), and 90th percentile absolute error (p90 error). For abstention rates > 14.7% the p90 error of the deep learning model is below 3.22 years and better compared to the human reader estimate (p90 error = 3.29 years). For abstention rates > 82.9% the max error of the deep learning model is below 3.49 years and better compared to the human reader estimate (max error = 3.78 years)

Table 4 Summary of radiological age assessment results for the test set using deep learning (DL) and optimistic human reader performance estimate (HRE). The table displays the mean absolute error (MAE), absolute error standard deviation (SD), maximum absolute error (max error), and 90th percentile absolute error (p90 error) for female and male persons. Additionally, the table shows the deep learning results with an abstention-performance trade-off for abstention rates of 20% and 50%, where predictions with the highest predictive uncertainty (20 or 50% of predictions) are not taken into account for analysis

	MAE (SD)		max error		p90 error	
	f	m	f	m	f	m
HRE	1.84 (0.84)		3.40	3.78	3.29	
	1.77 (0.74)	1.91 (0.92)			3.08	3.42
DL	1.65 (1.27)		6.40	7.32	3.38	
	1.69 (1.18)	1.62 (1.34)			3.10	3.45
DL (20% AR)	1.57 (1.14)		5.44	5.32	3.12	
	1.61 (1.09)	1.52 (1.18)			3.06	3.36
DL (50% AR)	1.35 (1.00)		4.24	5.32	2.67	
	1.44 (0.99)	1.26 (1.00)			2.91	2.54

Fig. 6. All metrics decreased for increasing abstention rates, i.e. the greater the fraction of the most uncertain predictions that were not considered for analysis, the better the remaining predictions on average. For abstention rates > 14.7% the p90 error of the deep learning model was below 3.22 years and outperformed the human reader estimate which had a p90 error of 3.29 years. For abstention rates > 82.9% the max error of the deep learning model was below 3.49 years and thus better compared to the human reader estimate which had a max error of 3.78 years. Table 4 reports the deep learning model's and human reader estimate's MAE, max error, and p90 error separately for female and male individuals, and for abstention rates of 20% and 50%.

Discussion

Radiological age assessment methods based on reference studies analyzing the ossification of the sterno-clavicular joint are inherently limited in accuracy due to their design. The clavicles specifically enable age assessment for older minors (15–18 years), adolescents (18–21 years), and young adults (21–30 years). In an optimistic human reader performance estimate, we calculated that the well-established method of Kellinghaus et al. [7–9] cannot predict chronological age more accurately than 1.84 years on average and no better than 0.66 years at best for individuals whose true age is between 15.0 and 30.0 years.

Deep learning model

In an effort to overcome this inherent limitation, we developed a deep learning approach for radiological age assessment based on clavicle ossification (Fig. 2). The deep learning model outperformed the human reader estimate of the Kellinghaus et al. method on average and achieved a MAE of 1.65 years on a balanced test dataset containing 300 chest CT volumes that have been cropped around the sterno-clavicular joints.

While the superior average performance highlights the potential of deep learning, ensuring the algorithm's safety for all individuals is crucial. Consequently, the model's highest error should be low and high errors should be infrequent during testing. Deep learning returned absolute errors up to 7.32 years and fell short of the human reader estimate which only had absolute errors up to 3.78 years (Table 4). However, the samples that returned the worst deep learning predictions showed norm-variants or pathologic disorders, which would be exclusion criteria for radiological age assessment with reference study methods [23, 24].

Additionally, rare high error predictions can be avoided for deep learning with an abstention-performance trade-off (Fig. 6): we leveraged predictive uncertainty to identify potential high error predictions, excluded them from analysis, and improved performance for the remaining predictions. For abstention rates > 14.7% the deep learning model surpassed the human reader estimate's p90 error of 3.29 years, indicating the potential for reducing high errors during application.

Another benefit of automated deep learning age assessment is the significantly reduced analysis time for scans. This advantage may valuable in post-mortem CT examinations for identification purposes, e.g. following mass casualty incidents.

Positioning within the literature

To the best of our knowledge, no deep learning-based age assessment using chest CT volumes of the clavicles has been reported yet. However, several pioneering studies leverage other imaging modalities to predict age based on different skeleton areas. Auf der Mauer et al. [14] analyzed 185

coronal and 404 sagittal 3D knee MRI volumes of Caucasian male subjects between the age of 13.0 and 21.8 years and middle to high socio-economic status. Using a combination of a deep learning model and a classical decision tree-based machine learning algorithm, they could improve the MAE from 1.63 (SD = 0.99) years achieved by a naive baseline model, which always predicts the mean age of the training set, to 0.69 (SD = 0.49) years. Vila-Blanco et al. [13] studied 2,289 2D dental panoramic radiograph images of Spanish Caucasian subjects in the age range of 4.5 to 89.2 years. Their deep learning model achieved an MAE of ~ 2.5 years for the subgroup of 798 subjects between the ages of 15.0 and 30.0 years. In the 2017 RSNA Pediatric Bone Age Machine Learning Challenge [25], participants trained deep learning models to predict expert-assigned bone age.

Limitations

This study has limitations. First, the complex relationship between skeletal development and chronological age poses an insurmountable natural accuracy barrier [26] for age assessment and depends on a variety of factors ranging from genetic predisposition to socio-economic status [27]. Second, the data used to train, validate, and test the deep learning model was collected retrospectively and acquired during the clinical routine for all purposes. Therefore, it was inhomogeneous, acquired with different scanners using different protocols, and includes samples that would have been ruled out for radiological age estimation by experts based on the health condition of the individual. Third, all CT scans in our dataset were acquired at the same hospital, which likely introduced a bias that prevents the data from being representative of the global population. Fourth, the training dataset included multiple CT scans per individual (4400 CT scans vs. 1935 individuals), while only one unique scan per individual was used in the validation and test dataset (300 CT scans and 300 individuals, respectively). Additionally, the dataset included only CT scans for which the automated localization of the medial clavicular epiphyseal cartilages returned a positive detection. Finally, the human reader estimate is based on the statistics reported by Kellinghaus et al. [7, 8], but other studies applying the same method exist, e.g. from Wittschieber et al. [9].

Ethics disclaimer

We do not endorse the actual or exploratory application of the approach presented in this study for radiological age assessment. Instead, we suggest further research into deep learning approaches for radiological age assessment under controlled settings, following the promising results in this and similar studies. Specifically, we recommend transferring the presented approach from CT to magnetic resonance imagining (MRI) data to avoid exposing individuals to potentially harmful ionizing radiation. The MRI dataset should be extensive and inclusive to ensure its representation of all individuals. The research should also focus on reducing deep learning prediction variance and extreme errors.

Conclusion

In summary, our study demonstrates a deep learning approach for radiological age assessment using CT volumes that highlight the medial clavicular epiphyseal cartilages. Deep learning surpassed the human reader performance estimate in terms of mean accuracy (MAE = 1.65 vs. 1.84 years). Errors could partially be attributed to physiological abnormalities. Also, high errors may be avoided by abstaining from predictions with high uncertainty. Looking ahead, deep learning offers an accurate, objective, and scalable solution that eliminates intra- and inter-reader variability and could be further improved with larger and standardized datasets.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-024-03167-6.

Acknowledgements Johanna Topalis was supported through funding from Bayerisches Staatsministerium für Wissenschaft und Kunst in cooperation with Fonds de Recherche Santé Québec.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors did not receive support from any organization for the submitted work.

Data availability The datasets generated and analyzed during the current study are not publicly available due to them containing information that could compromise research participant privacy, but are available from the corresponding author on reasonable request and with permission of the institutional review board (Ethics Committee, Medical Faculty, LMU Munich).

Code availability The underlying code for this work is available in *Github* and can be accessed via this link https://github.com/pwesp/dl-rad-age.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Informed consent The requirement for written informed consent was waived by the institutional review board (Ethics Committee, Medical Faculty, LMU Munich).

Ethical approval This retrospective study was approved by the institutional review board (Approval number 20-324, Ethics Committee, Medical Faculty, LMU Munich). All methods were carried out in accordance with the Declaration of Helsinki.

References

1. Schmeling A, Dettmeyer R, Rudolf E et al (2016) Forensic age estimation: Methods, certainty, and the law. Dtsch Arzteblatt Int 113:44–50. https://doi.org/10.3238/arztebl.2016.0044
2. Greulich WW, Pyle SI (1959) Radiographic Atlas of Skeletal Development of the Hand and Wrist. Stanford University Press, Stanford, CA, USA
3. European Asylum Support Office (2018) EASO Practical guide on age assessment. Publications Office
4. United Nations (1989) The Convention on the Rights of the Child. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child. Accessed 28 Feb 2023
5. The European Parliament and The Council Of The European Union (2013) Directive 2013/33/EU of the European Parliament and of the Council of 26 June 2013 laying down standards for the reception of applicants for international protection (recast)
6. Schmeling A, Schulz R, Reisinger W et al (2004) Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med 118:5–8. https://doi.org/10.1007/s00414-003-0404-5
7. Kellinghaus M, Schulz R, Vieth V et al (2010) Forensic age estimation in living subjects based on the ossification status of the medial clavicular epiphysis as revealed by thin-slice multidetector computed tomography. Int J Legal Med 124:149–154. https://doi.org/10.1007/s00414-009-0398-8
8. Kellinghaus M, Schulz R, Vieth V et al (2010) Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans. Int J Legal Med 124:321–325. https://doi.org/10.1007/s00414-010-0448-2
9. Wittschieber D, Schulz R, Vieth V et al (2014) The value of substages and thin slices for the assessment of the medial clavicular epiphysis: a prospective multi-center CT study. Forensic Sci Med Pathol 10:163–169. https://doi.org/10.1007/s12024-013-9511-x
10. Thodberg HH (2009) An Automated Method for Determination of Bone Age. J Clin Endocrinol Metab 94:2239–2244. https://doi.org/10.1210/jc.2008-2474
11. Tajmir SH, Lee H, Shailam R et al (2019) Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. Skeletal Radiol 48:275–283. https://doi.org/10.1007/s00256-018-3033-2
12. Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
13. Vila-Blanco N, Carreira MJ, Varas-Quintana P et al (2020) Deep neural networks for chronological age estimation from OPG images. IEEE Trans Med Imaging 39:2374–2384. https://doi.org/10.1109/TMI.2020.2968765
14. auf der Mauer M, Jopp-van Well E, Herrmann J, et al (2021) Automated age estimation of young individuals based on 3D knee MRI using deep learning. Int J Legal Med 135:649–663. https://doi.org/10.1007/s00414-020-02465-z
15. Dallora AL, Anderberg P, Kvist O et al (2019) Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. PLoS One 14:e0220242. https://doi.org/10.1371/journal.pone.0220242
16. Chartrand G, Cheng PM, Vorontsov E, et al (2017) Deep learning: A primer for radiologists. Radiogr Rev Publ Radiol Soc N Am Inc 37:2113–2131. https://doi.org/10.1148/rg.2017170077
17. Cunningham C, Scheuer L, Black S (2016) Developmental juvenile osteology. Academic press
18. Levin E, Tishby N, Solla SA (1990) A statistical approach to learning and generalization in layered neural networks. In: Proceedings of the IEEE. pp 1568–1574
19. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp 770–778
20. Wesp P, Sabel BO, Mittermeier A, et al (2023) Automated localization of the medial clavicular epiphyseal cartilages using an object detection network: a step towards deep learning-based forensic age assessment. Int J Legal Medhttps://doi.org/10.1007/s00414-023-02958-7
21. Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc., Red Hook, NY, USA, Long Beach, CA, USA
22. Wesp P, Schachtner BM, Grosu S et al (2022) Allowing machine learning models to say "I don't know": Improving automated clinical decision-making by balancing performance against abstention. European Society of Radiology, Vienna, Austria, Vienna, Austria
23. Rudolf E, Kramer J, Schmidt S et al (2019) Anatomic shape variants of extremitas sternalis claviculae as collected from sternoclavicular thin-slice CT-studies of 2820 male borderline-adults. Int J Legal Med 133:1517–1528. https://doi.org/10.1007/s00414-019-02065-6
24. De Tobel J, Hillewig E, van Wijk M et al (2020) Staging Clavicular Development on MRI: Pitfalls and Suggestions for Age Estimation. J Magn Reson Imaging 51:377–388. https://doi.org/10.1002/jmri.26889
25. Halabi SS, Prevedello LM, Kalpathy-Cramer J et al (2019) The RSNA pediatric bone age machine learning challenge. Radiology 290:498–503. https://doi.org/10.1148/radiol.2018180736
26. Dahlberg PS, Mosdøl A, Ding Y et al (2019) A systematic review of the agreement between chronological age and skeletal age based on the Greulich and Pyle atlas. Eur Radiol 29:2936–2948. https://doi.org/10.1007/s00330-018-5718-2
27. Schmeling A, Reisinger W, Loreck D et al (2000) Effects of ethnicity on skeletal maturation: Consequences for forensic age estimations. Int J Legal Med 113:253–258. https://doi.org/10.1007/s004149900102
28. Paszke A, Gross S, Massa F, et al (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in neural information processing systems. Curran Associates, Vancouver, Canada
29. Kingma DP, Ba JL (2015) Adam: A Method for Stochastic Optimization. In: 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA

List of Figures

List of Tables

Danksagung

Michi, mein erster Dank gehört dir. Nicht nur weil du mein Doktorvater bist und sich das, dem durchblättern zahlreicher Doktorarbeiten nach zu urteilen, so gehört, sondern weil ich wirklich dankbar bin für deine Betreuung. Ich bin dankbar für das große Vertrauen und die Freiheit bei meiner Arbeit und deine lockere, immer ehrliche Art. Genauso dankbar bin ich für deinen unermüdlichen Kampf gegen die Bürokratie, der deiner Gruppe den Rücken freihält, und dafür, dass du ihn jederzeit für uns unterbrichst, falls es mal Probleme gibt. Eine bessere Doktorandenzeit als bei dir hätte ich mir nicht wünschen können!

Andi, Balthasar und Moritz, euch Dreien möchte ich besonders danken. Ihr habt mich vor vier Jahren in Großhadern aufgenommen und mir nicht nur Paper schreiben, sondern auch Schafkopfen beigebracht. Ohne euch wärs nicht ansatzweise so lustig gewesen und ich bin froh, dass wir Kollegen und Freunde geworden sind.

Jakob, Johanna, Kathi, Olaf, Tessi, Timo und Tobi, dank euch hat mir die Arbeit immer Spaß gemacht und ich bin gern ins Büro gekommen wenn ich wusste, dass ihr da seid. Außerdem ist meine Doktorarbeit genau wie jedes Paper ein Gemeinschaftswerk und ohne eure Hilfe hätte ich sicher kein Projekt zu Ende gebracht.

Basti, Clemens, Sergio, danke, dass ihr mir immer mit Rat und Tat zur Seite standet, auch abseits radiologischer Fragen, und, dass ihr so viele spannende Forschungsrojekte an Land zieht. Ohne euch hätte es keines der Paper gegeben, die jetzt meine Doktorarbeit füllen. Sergio, vielen Dank für das Vertrauen in mich, meine Einstellung als Doktorand habe ich auch dir zu verdanken, und für die wirklich angenehme Zusammenarbeit seit Tag eins.

Prof. Ricke, auch Ihnen möchte ich danken, für das Interesse und die Aufgeschlossenheit gegenüber Data Science und Machine Learning in der Radiologie und die Unterstützung unserer Forschung.

George und Guillaume, den Weg von der Physik hin zu Machine Learning und medizinischer Forschung habe ich auch dank euch gefunden. George, du hast mich auf Michis Doktorandenstelle aufmerksam gemacht und wer weiß, ob es mir woanders so gut ergangen wäre wie hier, vielen Dank dafür.

Ruben, vielen Dank an den besten Physiklehrer der Welt und die Ermutigung zum Physikstudium. Ich hoffe du kannst weiterhin so viele Schüler für Naturwissenschaft begeistern.

Mama, Papa, euch möchte ich natürlich auch Danke sagen. Vielen Dank für eure bedingungslose Unterstützung bei allem was ich im Leben tue, einschließlich meinem Studium und der Doktorarbeit. Auch wenn ihr nicht immer genau wisst, was ich eigentlich den ganzen Tag mache, weiß ich, dass ihr stolz auf mich seid. Ich bin froh euer Sohn zu sein, dass wir uns so gut verstehen und ich immer auf euch zählen kann.

An meine Freunde aus Ladenburg und München, mit niemandem macht Klamauk so viel Spaß wie mit euch, und es ist immer schön mit euch gemeinsam die Arbeit Arbeit sein zu lassen, egal ob im Urlaub, beim Wandern, Radeln oder Skifahren.

Alli, seit dem ersten Semester bist du an meiner Seite und der wichtigste Mensch in meinem Leben. Danke für die letzten zehn Jahre mit dir, die nächsten zehn kann ich kaum erwarten.