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

Praktische Anwendung von Deep Learning: Klassifizierung der häufigsten Röntgenbilder in einem PACS mit Hilfe eines neuronalen Netzwerkes

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

vorgelegt von
Dr. phil. Thomas Markus Dratsch
aus Wuppertal

promoviert am 23. Oktober 2023

Gedruckt mit Genehmigung der Medizinischen Fakultät der Universität zu Köln
2023

Dekan: Universitätsprofessor Dr. med. G. R. Fink
1. Gutachter: Privatdozent Dr. med. D. Pinto dos Santos
2. Gutachter: Universitätsprofessor Dr. med. Dip.-Psych. J. Kambeitz

Erklärung

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

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

Herr PD Dr. med. Daniel Pinto dos Santos

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

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

Für die vorliegende kumulative Dissertation habe ich das Paper selbstständig verfasst. Außerdem habe ich selbstständig das neuronale Netzwerk trainiert und die Ergebnisse mit SPSS ausgewertet. Die zugrundeliegenden Röntgenbilder wurden in Zusammenarbeit mit PD Dr. med. Pinto dos Santos selbstständig sortiert.

Erklärung zur guten wissenschaftlichen Praxis:

Ich erkläre hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten (Amtliche Mitteilung der Universität zu Köln AM 132/2020) der Universität zu Köln gelesen habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen.

Köln, den 26.09.2022

Unterschrift: .....

Danksagung

Ich möchte mich bei allen Personen bedanken, die mich bei dieser Promotion begleitet haben. Besonderer Dank gilt Daniel Pinto dos Santos für die enge Betreuung während des Projekts. Außerdem möchte ich mich bei meinen Eltern für Ihre Unterstützung bedanken.

Inhaltsverzeichnis

Abkürzungsverzeichnis

AI = Artificial Intelligence (engl.: Künstliche Intelligenz)

DICOM = Digital Imaging and Communications in Medicine

JPEG = Joint Photographic Expert Group

PACS = Picture archiving and communication system

OSG = oberes Sprunggelenk

LWS = Lendenwirbelsäule

ap = anterior-posterior

MTA = medizinisch-technische/r Assistentin/Assistent

pa = posterior-anterior

CT = Computertomographie

MRT = Magnetresonanztomographie

PET = Positronen-Emissions-Tomographie

1. Zusammenfassung

Das Ziel dieser Arbeit war es, ein neuronales Netzwerk zur Klassifikation der häufigsten Kategorien von konventionellen Röntgenbildern (z.B.: Thorax ap, Abdomen in Seitenlage) zu entwickeln und anhand von internen und externen Daten zu validieren. Ein solches Netzwerk kann dabei helfen verschiedene radiologische Arbeitsabläufe zu verbessern. Hierzu wurden alle an unserem Institut erstellten Röntgenbilder aus dem Jahr 2017 (*n* = 71.274) aus dem PACS (Picture Archiving and Communication System) aufgerufen. Die 30 größten Kategorien (*n* = 58.291, 81,7% aller im Jahr 2017 erstellten Röntgenbilder) wurden dazu verwendet ein neuronales Netzwerk (MobileNet v1.0) mittels Transfer Learning zu trainieren und zu validieren. Die Kategorien der Röntgenbilder wurden anhand der DICOM-Metadaten extrahiert und an die Kategorien des WHO Manuals of Diagnostic Imaging angepasst. Zur unabhängigen, externen Validierung der Ergebnisse dienten Bilder von externen Krankenhäusern aus unserem PACS (*n* = 5324). In der internen Validierung betrug die Genauigkeit des Modells 90.3% (95%CI: 89.2–91.3%), In der externen Validierung betrug die Genauigkeit des Modells 94.0% (95%CI: 93.3–94.6%). Mit Hilfe von Daten nur einer Institution waren wir in der Lage ein neuronales Netzwerk zur Klassifikation der häufigsten Kategorien von Röntgenbildern zu trainieren. Das Netzwerk zeigte eine gute Generalisierbarkeit in den externen Daten und kann dazu verwendet werden Bilder deren Metadaten fehlen oder fehlerhaft sind in einem PACS zu organisieren bzw. eine Vorauswahl an Röntgenbildern zu treffen, so dass diese an spezialisierte neuronale Netzwerke zur Erkennung von Erkrankungen weitergeleitet werden können. Das neuronale Netzwerk kann auch dabei helfen andere radiologische Arbeitsabläufe zu optimieren (zum Beispiel: automatisierte Aufhängung von Röntgenbildern; Überprüfung, ob angefordertes und durchgeführtes Bild übereinstimmen). Das finale neuronale Netzwerk steht öffentlich zur Evaluation und Erweiterung zur Verfügung.

2. Einleitung

2.1. Einleitung

Öffnet man in jüngerer Zeit eine medizinische Fachzeitschrift, so ist die Wahrscheinlichkeit hoch auf einen Artikel zu stoßen, der den Einfluss von künstlicher Intelligenz auf die jeweilige Subdisziplin diskutiert. An vorderster Front befindet sich hierbei die Radiologie, welche durch ihre Nähe zur Technik besonders für eine Umwälzung durch neue Technologien prädestiniert zu sein scheint. Die Vorhersagen reichen hier von künstlicher Intelligenz, welche Radiologinnen und Radiologen bei Ihrer Arbeit unterstützt, bis zu Szenarien, in welchen Radiologinnen und Radiologen vollständig durch künstliche Intelligenz ersetzt werden[1,2]. Eines ist offensichtlich: Künstliche Intelligenz wird in den nächsten Jahren auf verschiedene Weise ihren Weg in den radiologischen Alltag finden.

Ein Großteil der aktuellen Forschung in diesem Gebiet konzentriert sich auf den Teil der Arbeit von Radiologinnen und Radiologen, der Wahrnehmung und logisches Schlussfolgern umfasst (z.B.: das Erkennen von Anomalitäten und Diagnostizieren von Erkrankungen)[3,4]. Während diese Ansätze auf den ersten Blick vielversprechend erscheinen, so gibt es doch verschiedene Hindernisse—ethisch, ökonomisch und juristisch—die eine Einführung in diese Bereiche der radiologischen Arbeit erschweren. Demgegenüber steht ein Ansatz—vertreten in dieser Arbeit—, dass künstliche Intelligenz vor allem als Werkzeug zur Qualitätssicherung (z.B.: Strahlenschutz oder das Erkennen von suboptimalen Aufnahmen) und bei der Automatisierung sich wiederholender, einfacher Alltagsaufgaben (z.B.: Messungen, Segmentierungen und die Erkennung von Fremdmaterialien) einen positiven Beitrag in der Radiologie leisten kann[5].

Ein Beispiel hierfür ist die korrekte Beschriftung von Bildern in einem PACS (Picture Archiving System), bei der alle relevanten Informationen über ein Bild (z.B.: Patientendaten, applizierte Dosis, und erfasste Körperregion) als Metadaten gemeinsam mit den Bilddaten im Dateiformat DICOM festgehalten werden. Diese Daten können jedoch inkorrekt sein, wie erstmals von Güld et al. gezeigt wurde. Diese fanden heraus, dass die Informationen unter dem Eintrag *erfasste Körperregion* in 15.3% inkorrekt waren[6]. Dies ist aus verschiedenen Gründen problematisch: Zum einen wird so die Erstellung von Datensätzen für Machine Learning Projekte erschwert, wenn nicht alle Bilder zu einer Körperregion gefunden werden können. Zum anderen wird so die Einrichtung von automatischen Arbeitsabläufen erschwert, bei denen Bilder je nach Körperregion an spezifische Klassifikationsalgorithmen weitergeleitet werden. Darüber hinaus bauen viele weitere radiologische Arbeitsaufläufe auf einer korrekten Beschriftung von Bildern auf: Ein neuronales Netzwerk, das die erfasste Körperregion in einem Röntgenbild korrekt erkennt, kann zum Beispiel helfen zu überprüfen, dass Untersuchungen nicht unnötig mehrfach angefordert werden und dass die erfolgte Untersuchung der

angeforderten Untersuchung entspricht. Zusätzlich kann ein solches Netzwerk dabei helfen automatische Aufhängungsprotokolle zu verbessern.
Konventionelle Röntgenbilder sind weiterhin die am meisten durchgeführte Bildgebung in der Radiologie. Das Ziel dieser Arbeit war es daher, ein neuronales Netzwerk zur Klassifikation der häufigsten konventionellen Röntgenbilder (z.B.: Thorax ap, Abdomen in Seitenlage) zu entwickeln und in einem externen Datensatz zu validieren. Das fertige Modell wird öffentlich zur Verfügung gestellt, so dass es auch unabhängig evaluiert und in radiologische Arbeitsabläufe integriert werden kann.
Die Struktur dieser Arbeit ist wie folgt: Zuerst wird eine kurze Einführung in die verschiedenen Unterarten von Machine Learning gegeben. Hierauf folgt ein Überblick über die verschiedenen Schritte des Trainierens und Validierens eines neuronalen Netzwerkes. Danach werden verschiedene Einsatzmöglichkeiten von neuronalen Netzwerken in der Medizin und Radiologie sowie deren Probleme diskutiert. Darauf folgt ein Überblick über den Einsatz von neuronalen Netzwerken als Werkzeug zur Qualitätssicherung und bei der Automatisierung sich wiederholender, einfacher Alltagsaufgaben. Zum Abschluss werden Einschränkungen der aktuellen Arbeit diskutiert und ein Ausblick auf zukünftige Weiterentwicklungen gegeben.

2.2. Geschichte von Machine Learning

Machine Learning wird im Allgemeinen als eine Subdisziplin des Forschungsbereiches der künstlichen Intelligenz aufgefasst, eine Disziplin, die sich mit der Nachbildung durch Maschinen von früher nur dem Menschen zugesagten kognitiven Fähigkeiten wie Denken und logischem Schlussfolgern beschäftigt. Während künstliche Intelligenz sich allgemein mit der Fähigkeit von Maschinen intelligentes Verhalten zu zeigen befasst, so konzentriert sich Machine Learning im engeren Sinne auf die Fähigkeit von Maschinen basierend auf Daten zu lernen—einer der Grundpfeiler für intelligentes Verhalten.

Vorstellungen von nicht-menschlichen Maschinen oder Wesen, die teilweise intelligentes Verhalten zeigen, gehen bis in die Antike zurück—zum Beispiel in der Form von Talos, einer lebendigen Bronzestatur, die selbstständig die Insel Kreta bewachte—, tauchen in der jüdischen Mythologie des Mittelalters auf—hier in der Form des Golems—und finden sich in der romantischen Literatur des 19. Jahrhunderts als Frankensteins Monster. Gepaart mit der während der Aufklärung aufkommenden Faszination mit Logik und Vernunft und der Vorstellung—vor allem von Leibniz vorangetrieben—dass sich alle Fragen der Wissenschaft mit Hilfe einer Logik der Vernunft berechnen lassen, mündeten diese Ideen in der Figur des Roboters—bekannt geworden durch die Bücher von Issac Asimov—einer Maschine, die logisch denkt und autonom handelt[7].

Die erste wissenschaftliche Forschung im Bereich der künstlichen Intelligenz begann in den Nachkriegsjahren des 20. Jhdts. Neuartige Rechenmaschinen hatten während des Krieges beeindruckende Erfolge im Bereich der Kryptographie erzielt und ließen so den alten Traum von der denkenden Maschine näher rücken. So formulierte Alan Turing 1950 in seinem einflussreichen Aufsatz *Computing Machinery and Intelligence*[8]: "*I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.*" Einer der wichtigen Gründungsmoment der KI-Forschung war das Dartmouth Summer Research Project on Artificial Intelligence im Jahre 1956, bei dem viele Pioniere des Gebiets, wie John McCarthy, Marvin Minsky und Claude Shannon, zusammenkamen[9]. Ein Höhepunkt dieses Treffens war das Programm Logic Theorist von Allen Newell und Herbert A. Simon, welches dazu in der Lage war, automatisch 38 Beweise für 52 der mathematischen Sätze aus der Principia Mathematica zu finden[10]. Weitere Meilensteine dieser Anfangsphase waren die Entwicklung des Perceptrons im Jahre 1957[11]—einer der ersten Klassifikationsalgorithmen, dessen Struktur sich an der Architektur von menschlichen Neuronen orientierte—und das Programm ELIZA—welches menschliche Kommunikation so realistisch nachahmen konnte, dass einige Benutzer nicht merkten, dass sie mit einem Programm interagierten[12]. Geprägt war diese Anfangsphase der KI-Forschung durch einen unerschütterlichen Optimismus. So behauptete Herbert A. Simon im Jahre 1965[13]: "*Machines will be capable, within twenty years, of doing*

any work a man can do." Dieser Optimismus führte dazu, dass große Summen an Forschungsgeldern in die KI-Forschung investiert wurden[7].

Diese erste Phase der KI-Forschung erstreckte sich über die 50er und 60er Jahre. Anfang der 70er Jahre setzte jedoch eine erste Ernüchterung ein. Zwar war es gelungen in einzelnen, eng umgrenzten Bereichen Fortschritte zu erzielen, jedoch waren die Ergebnisse noch weit von den anfänglichen, optimistischen Prognosen entfernt. Zusätzlich wurde immer deutlicher, dass die bisherigen Techniken nicht dazu geeignet waren, jemals komplexere Probleme lösen zu können. Dies führte dazu, dass ein Großteil der Forschungsgelder gestrichen wurden und das Feld der KI-Forschung in eine Sinnkrise stürzte—diese Phase wird auch gemeinhin als der erste Winter der KI-Forschung bezeichnet[14].

Zu Beginn der 80er erlebte die KI-Forschung erneut vermehrt Interesse. Zum Teil als Reaktion auf ein groß angelegtes Forschungsprojekt der japanischen Regierung im Jahre 1981 wurden auch im Westen wieder größere Summen in die KI-Forschung investiert. Im Gegensatz zur 1. Phase der KI-Forschung, in der das Ziel darin lag eine allgemeine künstliche Intelligenz zu schaffen, rückte nun das Nachbilden von Expertenwissen in den Fokus. Solche sogenannten Expertensysteme sollten in eng umschriebenen Bereichen—zum Beispiel der Analyse von Daten der Massenspektrometrie—Experten ergänzen oder ersetzen können. Eines der ersten Expertenysteme im Bereich der Medizin war MYCIN[15], welches zur Diagnostik und Therapie von Blutstrominfektionen eingesetzt wurde. Bis zum Ende der 80er-Jahre wurden zahlreiche Expertensystem entwickelt und in der Industrie eingesetzt. Der anfängliche Enthusiasmus für diese neue Technologie ebbte erneut ab, als sich mit der Zeit herausstellte, dass Expertensysteme aufwendig in der Unterhaltung und nur schwer weiterzuentwickeln waren. Wieder kam es zu einer erneuten Phase der Abnahme von Forschungsgeldern und Interesse im Bereich der KI-Forschung, welche vom Ende der 80er-Jahre bis zur Mitte der 90er-Jahre reichte—der sogenannte 2. Winter der KI-Forschung[15].

In der Mitte der 90er begann dann eine Serie von Durchbrüchen in Gebieten, in denen es bisher nicht gelungen war Fortschritte zu erzielen: Im Jahr 1997 verlor der Schachweltmeister Gary Kasparov gegen den Computer Deep Blue[16], im Jahr 2005 legte erstmals ein autonom gesteuertes Auto eine Strecke von 131 Meilen zurück[17] und im Jahr 2011 gewann der von IBM entwickelte Computer Watson gegen zwei Champions in der Quiz-Show-Jeopardy[18]. Darüber hinaus gewannen neuronale Netzwerke immer mehr an Bedeutung. Im Jahr 2012 erzielte das neuronale Netzwerk Alexnet die höchste Erkennungsleistung im Rahmen des Bilderklassifikationswettbewerbes Imagenet[19]. Obwohl die grundlegende Struktur von neuronalen Netzwerken bis in die 50er Jahre zurück geht und die mathematischen Grundlagen für das Training dieser Netzwerke bereits in den 60er Jahren beschrieben wurde, fehlte es bisher an einer entsprechenden Menge an strukturierten Daten und Rechenleistung um neuronale Netzwerke effizient zu trainieren. Vor allem Fortschritte im Bereich von

leistungsstarken Graphikkarten machten nun das Training von neuronalen Netzwerken einer breiten Masse zugänglich, so dass diese sich zu einer der vorherrschenden Techniken im Bereich Machine Learning entwickeln konnten[20].

2.3. Struktur und Training eines neuronalen Netzwerkes

2.3.1. Grundlegende Struktur und Technik neuronaler Netzwerke

Die grundlegende Struktur von neuronalen Netzwerken ist den Verbindungen von Nervenzellen im menschlichen Kortex nachempfunden[21]. Eine Nervenzelle bzw. ein digitales Neuron enthält dabei Input von verschiedenen anderen Nervenzellen und gibt Impulse an nachfolgende Nervenzellen weiter. Die Verbindungen zwischen Nervenzellen können dabei unterschiedliche Gewichtungen haben. Im Rahmen des Trainings eines neuronalen Netzwerkes werden in einem iterativen Prozess diese Gewichtungen so lange adjustiert, bis das neuronale Netzwerk bei einem bestimmten Input (z.B.: dem Bild einer Blume) die korrekte Klassifikation ausgibt.

In einer einfachen Form besteht ein neuronales Netzwerk aus einem Input-Layer, einem Output-Layer sowie einer—je nach Komplexität der Klassifikationsaufgabe—mehr oder weniger großen Anzahl von Hidden-Layers zwischen Input- und Output-Layer (siehe Abbildung 1).

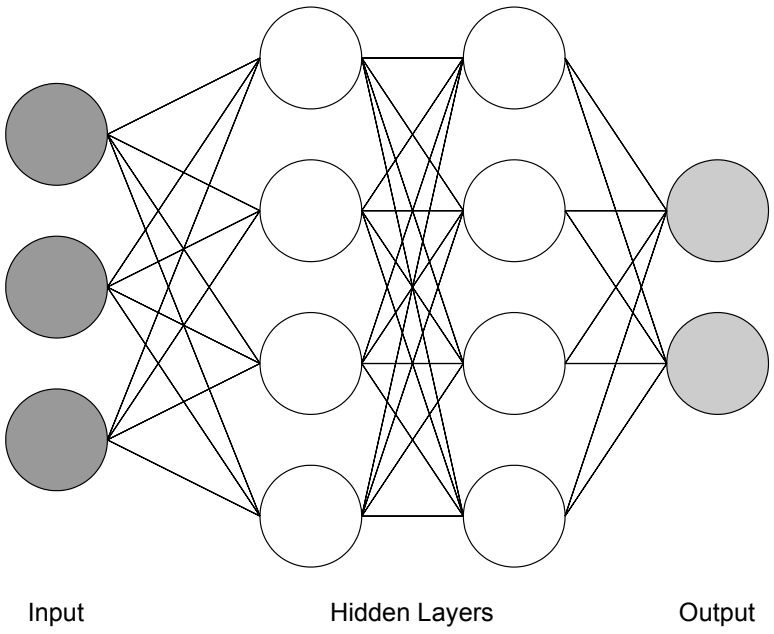

Abbildung 1: Struktur eines neuronalen Netzwerks.

Da ein Haupteinsatzgebiet von neuronalen Netzwerken das Klassifizieren von Bildern ist, soll die grundlegenge Struktur anhand eines Beispielnetzwerkes zur Klassifikation von handgeschriebenen Zahlen erläutert werden. Dieses Netzwerk soll in der Lage dazu sein, Bilder in einem Format von 28 x 28 Pixeln der Zahlen 0–9 korrekt zu klassifizieren. In diesem Fall besteht der Input-Layer aus 784 Neuronen, die den 784 Pixeln (28 x 28) der Bilder entsprechen, die das Netzwerk als Input erhält. Der Output-Layer besteht aus 10 Neuronen,

die den Zahlen 0–9 entsprechen, die das Netzwerk erkennen soll. Dazwischen liegt eine der Komplexität der Aufgabe entsprechende Anzahl an Hidden Layers.

2.3.2. Datenaufbereitung

Um ein neuronales Netzwerk zu trainieren, ist eine bestimmte Menge an strukturierten Daten nötig—in unserem Beispiel sind dies Bilder von handgeschriebenen Zahlen und die dazugehörigen korrekten Zahlen. Bei der Auswahl der Daten ist es wichtig, dass die im Training verwendeten Daten den Daten ähnlich sind, die das neuronale Netzwerk in Zukunft im Rahmen seines Einsatzes erhält. Ist dies nicht der Fall, so kann es passieren, dass das neuronale Netzwerk zwar gute Klassifikationsleistungen im Rahmen des Trainings erzielt, jedoch diese nicht in den eigentlichen Einsatzbereich generalisieren. Um das Netzwerk zu trainieren, werden die Daten aufgeteilt. Ein Teil der Daten wird zum Training verwendet und der Rest der Daten zur Validierung des trainierten Netzwerkes. Hierbei haben sich unterschiedliche Aufteilungsmethoden etabliert. Gemeinhin wird ein Verhältnis von 80 zu 20 oder 90 zu 10 verwendet. Dies bedeutet, dass ca. 80 % der Daten für das Training und 20% der Daten für die Validierung verwendet werden.

Um die Robustheit des neuronalen Netzwerkes zu erhöhen—das heißt die Fähigkeit eine große Bandbreite an neuen Daten korrekt zu klassifizieren—können die Daten im Rahmen des Trainingsprozesses modifiziert werden—im Bereich des Machine Learnings wird dieser Prozess auch als Augmentierung bezeichnet. Im Falle unseres Netzwerkes zur Erkennung von handgeschriebenen Zahlen könnte dies bedeuten, dass ein Teil der Bilder vor dem Training automatisch rotiert wird, um es dem Netzwerk zu ermöglichen auch Zahlen korrekt zu erkennen, die nicht streng gerade geschrieben sind. Sind die Daten aufbereitet, korrekt klassifiziert und gegebenenfalls augmentiert worden, so kann mit dem Training begonnen werden.

2.3.3. Training

Wie bereits beschrieben, ist es das Ziel des Trainings kontinuierlich die Gewichte des Netzwerkes zu optimieren, so dass Input-Bilder von dem Netzwerk korrekt klassifiziert werden. Als Ausgangszustand werden hierbei meistens randomisierte Gewichte verwendet. Zu Beginn des Trainings werden nun Bilder in das Netzwerk eingelesen und das Netzwerk generiert eine Klassifikation. Diese Klassifikation wird nun mit der wirklichen Klasse des Input-Bildes verglichen. Die Abweichung von der wirklichen Klassifikation und der von dem neuronalen Netzwerk generierten Vorhersage wird als Fehler bezeichnet. Hierauf werden die Gewichte des Netzwerkes adjustiert und der Prozess mit weitern Input-Bildern wiederholt. Zur Verbesserung des Netzwerkes wird hierbei der Gradient Descent-Algorithmus verwendet, der dabei hilft, die Gewichte des Netzwerkes zu optimieren. Wenn ein Netzwerk erfolgreich lernt,

so nimmt mit dem Verlauf des Trainings der Fehler ab und die korrekte Klassifikationsleistung des Netzwerkes zu. Dies geschieht bis zu einem bestimmten Punkt, um den die Klassifikationsleistung herum stagniert. Die maximale Klassifikationsleistung ist dann erreicht und das Training beendet.

2.3.4. Validierung

Um die Klassifikationsleistung des trainierten Modells zu bewerten, werden nun die Validierungsdaten verwendet, die vor dem Training vom Trainingsdatensatz separiert wurden und die das neuronale Netzwerk während des Trainings nicht gesehen hat. Die Validierungsdaten ermöglichen eine Abschätzung, wie gut das Netzwerk neue, unbekannte Daten klassifizieren kann. Die Validierungsdaten werden nun vom Netzwerk klassifiziert und die Vorhersagen mit den eigentlichen Klassen verglichen. Um die Klassifizierungsleistung des Netzwerkes zu quantifizieren, können verschiedene Parameter wie Genauigkeit (Accuracy), Sensitivität, Spezifität, positiver prädiktiver Wert und negativer prädiktiver Wert berechnet werden.

2.4. Neuronale Netzwerke in der Radiologie

2.4.1. Neuronale Netzwerke zur Klassifikation von Röntgenbildern

Als medizinische Disziplin, die vornehmlich mit Bildern arbeitet, scheint die Radiologie prädestiniert für den Einsatz neuronaler Netzwerke zur Diagnostik von Erkrankungen. Nachdem neuronale Netzwerke die beste Leistung im Bilderklassifikationswettbewerb ImageNet erreicht hatten[19], dauert es nicht lange bis zu den ersten Anwendungen in der Radiologie. Eines der ersten großen Projekt in der Radiologie, welches neuronale Netzwerke nutzte, ist das Netzwerk CheXNet[22]. Dieses ist dazu in der Lage 14 verschiedene Pathologien (z.B.: Pneumonie, Pleuraerguss oder Pneumothorax) in Thorax-Röntgenbildern zu erkennen. Neben Röntgenbildern des Thorax wurden neuronale Netzwerke dazu trainiert, Pathologien in Röntgenbildern des Abdomens zu erkennen. So trainierten zum Beispiel Cheng et al. ein neuronales Netzwerk eine Darmpassagestörung in Röntgenbildern des Abdomens zu erkennen[23]. Auch zur Klassifikation und Analyse von Röntgenbildern des Skelettsystems sind neuronale Netzwerke bereits eingesetzt worden. Einige der Anwendungen umfassen das Erkennen von Frakturen des Handgelenks[24] und der Hüfte[25] sowie das Diagnostizieren von arthrotischen Veränderungen des Kniegelenks[26].

2.4.2. Neuronale Netzwerke in der Schnittbildgebung

Der Einsatz von neuronalen Netzwerken im Rahmen von Schnittbildgebung ist im Gegensatz zu konventionellen Röntgenbildern komplexer, da hier nicht nur ein einzelnes Bild analysiert werden muss, sondern ein dreidimensionaler Bilddatensatz. Zum Beispiel trainierten Kang et al. ein dreidimensionales neuronales Netzwerk zum Erkennen von suspekten Lungenrundherden in CT-Datensätzen der Lunge[27]. Oftmals werden bei der Analyse von Schnittbildgebung jedoch nur zweidimensionale neuronale Netzwerke verwendet, wie zum Beispiel in einer Studie von Tomita et al. zur Klassifikation von Wirbelsäulenfrakturen[28].

2.4.3. Probleme beim Einsatz von Neuronalen Netzwerken in der Radiologie

Obwohl neuronale Netzwerke dazu in der Lage sind bestimmte Pathologien zu erkennen, so stehen ihrem breiten Einsatz in der Radiologie bestimmte Hürden gegenüber. Soll ein neuronales Netzwerk trainiert werden, so muss zuerst sichergestellt werden, dass eine entsprechende Menge an Daten zur Verfügung steht, um ein neuronales Netzwerk mit einer ausreichenden Performance zu trainieren. Auch muss geklärt werden, ob die entsprechenden rechtlichen Grundlagen erfüllt sind, bereits vorhandene Daten von Patienten zu verwenden. Ist das neuronale Netzwerk erfolgreich trainiert worden, muss ausführlich getestet werden, wie robust das Netzwerk arbeitet.

Ist das neuronale Netzwerk anfällig gegenüber den Geräten, mit denen die Bilder aufgenommen wurden? Idealerweise sollte ein neuronales Netzwerk mit Bildern von verschiedenen Geräten trainiert werden, um sicherzustellen, dass das Netzwerk gegenüber

Variationen der Aufnahmegeräte robust ist[29]. Besonders problematisch ist es, wenn ein systematischer Zusammenhang zwischen dem Aufnahmegerät und den zu detektierenden Pathologien besteht. Dies ist zum Beispiel der Fall, wenn ein CT-Gerät eines Herstellers in der Notaufnahme steht und ein CT-Gerät eines anderen Herstellers hauptsächlich für die Versorgung ambulanter Patienten eingesetzt wird. Hier besteht die Gefahr, dass das neuronale Netzwerk Bildinformationen, die mit der Art des Aufnahmegerätes zusammenhängen, für die Klassifikation von Pathologien benutzt. Werden zum Beispiel Patienten mit einer Lungenembolie vornehmlich an dem CT-Gerät in der Notaufnahme untersucht, so besteht die Gefahr, dass das neuronale Netzwerk Bildinformationen, die spezifisch für dieses Gerät sind, für die Klassifikation verwendet. Dies kann letztendlich dazu führen, dass die Wahrscheinlichkeit einer Lungenembolie bei Patienten, die an dem CT-Gerät in der Notaufnahme untersucht werden, überschätzt und bei Patienten, die an dem ambulanten CT-Gerät untersucht werden, unterschätzt wird.

Sind die Entscheidungen des neuronalen Netzwerkes frei von Bias? Die Zusammenstellung des Trainingsdatensatzes eines neuronalen Netzwerkes kann dessen Entscheidungen beeinflussen. So haben Studien gezeigt, dass neuronale Netzwerke bei ihrer Entscheidungsfindung Informationen wie zum Beispiel die ethnische Herkunft, das Alter oder das Geschlecht von Patienten verwenden[30,31]. Ist der Trainingsdatensatz in Bezug auf diese Patientenmerkmale nicht ausgeglichen oder bestehen systematische Zusammenhänge zwischen diesen Patientenmerkmalen und bestimmten Pathologin, so kann dies dazu führen, dass das neuronale Netzwerk Entscheidungen mit einem Bias trifft. So konnten zum Beispiel Larrazabal et al. zeigen, dass ein neuronales Netzwerk zum Erkennen von Pathologien in Röntgen-Thorax-Bildern schlechtere Ergebnisse für Röntgenbilder von Frauen liefert, wenn der Trainingsdatensatz nicht genug Röntgenbilder von Frauen enthielt[32].

Wie geht das neuronale Netzwerk mit Bildern um, die sich deutlich von den Bildern im Trainigsdatensatz unterscheiden (Outlier)? Im Rahmen des klinischen Alltags kann es vorkommen, dass dem neuronalen Netzwerk Bilder präsentiert werden, die sich deutlich von den Bildern im Trainingsdatensatz unterscheiden (z. B.: seltene Fremdkörper, postoperativer Status oder anatomische Normvarianten). In diesen Fällen ist es wichtig, sicherzustellen, dass das neuronale Netzwerk bei unauffälligen Befunden nicht fälschlicherweise Pathologien diagnostiziert oder für den Patienten bedrohliche Befunde als unauffällig klassifiziert.

Bleibt die Performance des neuronalen Netzwerkes über die Zeit hinweg stabil? Wie bereits erwähnt, können multiple Faktoren die Leistung eines neuronalen Netzwerkes beeinflussen. So können Änderungen in Bildakquiseprotokollen, der Geräte sowie des Patientenkollektivs dazu führen, dass die Leistung eines neuronalen Netzwerkes mit der Zeit abnimmt. Aus diesem Grund ist es wichtig, die Leistung des neuronalen Netzwerkes kontinuierlich zu überwachen, um etwaigen Schwankungen entgegenzuwirken.

Führt die Einführung eines neuronalen Netzwerkes dazu, dass sich die Leistung von Radiologinnen und Radiologen verbessert? Idealerweise unterstützt ein neuronales Netzwerk Radiologen und Radiologen bei der Arbeit und führt zu einem besseren Ergebnis: In den Fällen, in denen das neuronale Netzwerk eine korrekte Diagnose liefert, folgen Radiologinnen und Radiologen dieser. In den Fällen, in denen das neuronale Netzwerk keine korrekte Diagnose liefert, verlassen sich die Radiologen und Radiologen auf ihr eigenes Urteil. Zahlreiche Studien aus dem Bereich der Mensch-Maschine-Interaktion konnten jedoch zeigen, dass Menschen sich übermäßig auf die Empfehlungen von Computersystemen verlassen und mit der Zeit auch falschen Empfehlungen folgen—ein Phänomen, das als Automation Bias bezeichnet wird. So fanden Glaube et al. heraus, dass Ärztinnen und Ärzte auch den falschen Empfehlungen eines neuronalen Netzwerkes zur Diagnose von Pathologien in Röntgen-Thorax-Bildern folgten[33]. Aus diesem Grund ist es wichtig, bei den potentiellen Nutzerinnen und Nutzern eines neuronalen Netzwerk einen kritischen Umgang, der offen die Stärken und Schwächen des neuronalen Netzwerkes thematisiert, zu kultivieren, um Automation Bias aktiv entgegen zu wirken.

Neuronale Netzwerke, die den interpretativen Anteil der radiologischen Arbeit übernehmen sollen (z. B.: das Diagnostizieren von Erkrankungen), stehen vor großen Herausforderungen. Das Hauptproblem ist hierbei die große Anzahl an Erkrankungen, die in einem radiologischen Bild potenziell vorhanden sein können. So sind die bisher entwickelten neuronalen Netzwerke nur dazu in der Lage eine sehr kleine Zahl von Pathologien zu erkennen. Dies führt für den befundenden Radiologen oder die befundende Radiologin zu keiner Zeitersparnis, da das entsprechende Bild immer noch nach Pathologien untersucht werden muss, die von dem Netzwerk nicht erkannt werden können. Neuronale Netzwerke, die sich auf den nicht-interpretativen Anteil der radiologischen Arbeit fokussieren—wie das in dieser Arbeit entwickelte Netzwerk—, versuchen im Gegensatz hierzu ein eng umschriebenes Problem zu lösen, was potenziell zu einer relevanten Zeitersparnis für Radiologinnen und Radiologen führen kann.

2.5. Praktische Anwendung von neuronalen Netzwerken in der Radiologie

2.5.1. Messungen

Objektive und reliable Messungen spielen in fast allen Bereichen der Radiologie eine bedeutende Rolle. So entscheidet zum Beispiel im Bereich der onkologischen Bildgebung die Tumorgröße über das Tumorstadium und so auch mit über die Therapie. Auch die Dynamik von Messungen über die Zeit ist wichtig für die Entscheidung ob beispielsweise ein Lymphknoten oder Lungenrundherd suspekt ist oder ob eine Therapie anschlägt. Obwohl die meisten Messungen in der Radiologie relativ einfach durchzuführen sind, so können sie doch einen nicht unerheblichen Anteil der Befundungszeit in Anspruch nehmen und stellen einen oftmals repetitiven Teil der radiologischen Arbeit dar. Ein Bereich der Radiologie in dem Messungen eine große Rolle spielen und der daher von einer Automatisierung durch neuronale Netzwerke profitieren kann ist die muskuloskelettale Radiologie. Erste neuronale Netzwerke zur Messung des medialen Femurtibialwinkels sowie des mechanisch–anatomischen Femurwinkles[34], des Risser-Stadiums[35] und des Knochenalters[36] haben das Potential auf verschiedenen Ebenen zu einer Qualitätsverbesserung führen: So kann durch automatische Messungen der Radiologe oder die Radiologin zeitlich und kognitiv entlastet werden, so dass mehr Zeit und kognitive Ressourcen für die Befundung zur Verfügung stehen. Darüber hinaus können automatische Messungen—wenn sie korrekt und reliabel sind—die Vergleichbarkeit zwischen verschiedenen Befunden erhöhen, da die Messungen immer von demselben System durchgeführt werden.

2.5.2. Erkennung von Fremdmaterialien

Das Beschreiben von Fremdmaterialien und die Bewertung von deren korrekter Lage ist ebenfalls Bestandteil der radiologischen Routine. Auch hier können neuronale Netzwerke zu Zeitersparnis führen und zur Qualitätskontrolle beitragen. Ein wichtiger Bereich hier sind Fremdmaterialien in Röntgen-Thorax-Bildern (z. B.: ein zentralvenöser Katheter oder ein Endotrachealtubus). Erste Anwendungen hier sind neuronale Netzwerke, die die Art des Fremdmaterials und die korrekte Lage ermitteln können[37].

Ein weiteres Problem, das neuronale Netzwerke lösen können, ist die Identifikation von Implantaten bei fehlender Dokumentation oder fehlendem Implantatpass. Ein erste Lösung hierfür ist ein neuronales Netzwerk, das Implantatart und Hersteller von Knie- und Hüftprothesen korrekt erkennen kann[38].

2.5.3. Strahlenschutz und Qualitätskontrolle

Um die potenzielle Strahlenbelastung des Patienten zu minimieren und eine diagnostische Aufnahme zu gewährleisten, muss vor jeder CT-Aufnahme zuerst der Scanbereich festgelegt werden. Dieser Scanbereich sollte so gewählt sein, dass alle für die Fragestellung relevanten Strukturen abgebildet werden und nicht relevante Strukturen keiner unnötigen Strahlung

ausgesetzt werden. Um dieses Problem zu lösen, trainierten Demircioglu et al. ein neuronales Netzwerk für CT-Aufnahmen des Thorax[39]. Dieses Netzwerk ist dazu in der Lage, vergleichbare Scanbereiche wie Radiologinnen und Radiologen festzulegen. Darüber hinaus waren diese Scanbereiche kleiner und somit strahlenschonender als die von medizinisch-technischen Assistenten ausgewählten Scanbereichen. Somit hat das Netzwerk das Potential, die Strahlenbelastung von Patienten zu reduzieren und die Planungsphase von CTs zu verkürzen.

Ein weiteres Problem, das neuronale Netzwerke lösen können, sind nicht diagnostische Aufnahmen. Als Teil des radiologischen Alltags kommt es immer wieder zu radiologischen Aufnahmen, die von der Qualität her nicht dazu geeignet sind, die entsprechende Fragestellung zu beantworten. Schlimmstenfalls wird dies erst bei der Befundung der Bilder bemerkt, wenn der Patient bereits den CT-Scanner oder den Röntgenbereich verlassen hat. Neuronale Netzwerke, die bereits unmittelbar nach der Aufnahme die Qualität bewerten und gegebenenfalls eine neue Aufnahme vorschlagen, können hier Abhilfe schaffen. Ein erstes Netzwerk für diesen Einsatz wurde von Somasundaram entwickelt [40]. Dieses Netzwerk ist dazu in der Lage, inadequate Bilder der HWS von Kindern zu erkennen. Hierbei kommt es relativ häufig zu nicht diagnostischen Bildern, da Kinder auf Grund mangelnder Kooperation nicht so einfach vor dem Röntgengerät positioniert werden können. Dieses neuronale Netzwerk kann somit einen positiven Beitrag zur Qualitätskontrolle leisten.

2.5.4. Klassifikation von Röntgenbildern

Alle relevanten Informationen zu einem erfassten Röntgenbild werden als Teil der Metadaten zusammen mit dem Bild gespeichert. Hierzu gehört auch die erfasste Körperregion. Wie Güld et al. zeigen konnten, war die Information unter dem Eintrag *erfasste Körperregion* in 15.3% der Fälle im untersuchten Datensatz inkorrekt. Ein neuronales Netzwerk, das die erfasste Körperregion in einem Röntgenbild korrekt erkennt, kann zum Beispiel helfen zu überprüfen, dass Untersuchungen nicht unnötig mehrfach angefordert werden und dass die erfolgte Untersuchung der angeforderten Untersuchung entspricht. Zusätzlich kann ein solches Netzwerk dabei helfen automatische Aufhängungsprotokolle zu verbessern. Aus diesem Grund war es Ziel dieser Arbeit, ein neuronales Netzwerk zur Klassifikation der häufigsten Kategorien von konventionellen Röntgenbildern (z. B.: Thorax ap, Abdomen in Seitenlage) zu entwickeln und anhand von internen und externen Daten zu validieren, welches nun im Folgenden genauer dargestellt wird.

3. Publikation

European Radiology (2021) 31:1812–1818
https://doi.org/10.1007/s00330-020-07241-6

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Check for updates

Practical applications of deep learning: classifying the most common categories of plain radiographs in a PACS using a neural network

Thomas Dratsch[1] · Michael Korenkov[1] · David Zopfs[1] · Sebastian Brodehl[2] · Bettina Baessler[3] · Daniel Giese[1] · Sebastian Brinkmann[4] · David Maintz[1] · Daniel Pinto dos Santos[1]

Received: 21 July 2020 / Accepted: 28 August 2020 / Published online: 28 September 2020

Abstract

Objectives The goal of the present study was to classify the most common types of plain radiographs using a neural network and to validate the network's performance on internal and external data. Such a network could help improve various radiological workflows.

Methods All radiographs from the year 2017 ($n = 71{,}274$) acquired at our institution were retrieved from the PACS. The 30 largest categories ($n = 58{,}219$, 81.7% of all radiographs performed in 2017) were used to develop and validate a neural network (MobileNet v1.0) using transfer learning. Image categories were extracted from DICOM metadata (study and image description) and mapped to the WHO manual of diagnostic imaging. As an independent, external validation set, we used images from other institutions that had been stored in our PACS ($n = 5324$).

Results In the internal validation, the overall accuracy of the model was 90.3% (95%CI: 89.2–91.3%), whereas, for the external validation set, the overall accuracy was 94.0% (95%CI: 93.3–94.6%).

Conclusions Using data from one single institution, we were able to classify the most common categories of radiographs with a neural network. The network showed good generalizability on the external validation set and could be used to automatically organize a PACS, preselect radiographs so that they can be routed to more specialized networks for abnormality detection or help with other parts of the radiological workflow (e.g., automated hanging protocols; check if ordered image and performed image are the same). The final AI algorithm is publicly available for evaluation and extension.

Key Points

- *Data from one single institution can be used to train a neural network for the correct detection of the 30 most common categories of plain radiographs.*
- *The trained model achieved a high accuracy for the majority of categories and showed good generalizability to images from other institutions.*
- *The neural network is made publicly available and can be used to automatically organize a PACS or to preselect radiographs so that they can be routed to more specialized neural networks for abnormality detection.*

Keywords Machine learning · Radiography · Artificial intelligence

✉ Thomas Dratsch
t.dratsch@mac.comn

[1] Institute of Diagnostic and Interventional Radiology, University Hospital Cologne, Kerpener Str. 62, 50937 Cologne, Germany

[2] Institute of Computer Science, Johannes Gutenberg University Mainz, Mainz, Germany

[3] Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, Zürich, Switzerland

[4] Department of General, Visceral and Cancer Surgery, University Hospital Cologne, Cologne, Germany

Abbreviations

AI	Artificial intelligence
DICOM	Digital Imaging and Communications in Medicine
JPEG	Joint Photographic Expert Group
PACS	Picture archiving and communication system

Introduction

Machine learning is predicted to have a huge impact on the field of radiology [1], augmenting and assisting radiologists [2]. With new papers being published every week, one central

question remains: What can machine learning do for the average radiologist? Currently, the majority of research in radiology seems to be focused on applying machine learning to the parts of the imaging pipeline that involve perception and reasoning (e.g., detection, quantification, and diagnostic reasoning) [3, 4]. However, due to various barriers (e.g., ethical, economical, and legal), this approach, while promising, may not be the optimal starting point for introducing artificial intelligence into the radiological workflow. Instead, artificial intelligence could be used as a tool for quality assurance and help with automating simple but tedious task encountered in clinical routine [5]. For example, one common challenge in a picture archiving and communication system (PACS) is that images are often labeled incorrectly in the corresponding DICOM tag. The problem of unreliable DICOM information was first demonstrated by Güld et al who found that the DICOM tag *Body Part Examined* was incorrect in 15.3% of cases [6]. This is not only problematic for the retrieval of images for the purpose of creating datasets but also hinders the development of imaging pipelines in which images are automatically routed to specific classification algorithms. Besides, many other parts of the radiological workflow rely on correctly labeled images. Thus, a neural network that can correctly classify and tag images could be used to check that exams are not repeated unnecessarily, control that the acquired image is the same as the one that was ordered, and streamline hanging protocols for optimal reporting on images. Because plain radiographs are still the most common type of imaging performed, a network for the classification of plain radiographs can have a meaningful impact on the radiological workflow. Therefore, the main goal of our study was to develop and validate a convolutional network to classify the most common types of plain radiographs (e.g., thorax pa, abdomen lateral). The final model will be made publicly available so that it can be evaluated and integrated into the radiological workflow.

Materials and methods

Radiographs

All radiographs from the year 2017 ($N = 71{,}274$) performed at our institution were retrieved from the PACS and categorized into 102 categories based on their DICOM metadata (study, series, and image description) according to the WHO manual of diagnostic imaging [7]. Because some categories contained only a small number of images, we limited ourselves to the 30 largest categories ($n = 58{,}219$), which accounted for 81.7% (58,219/71274) of all radiographs performed in the year 2017 at our institution. For these 30 categories, all images were reviewed again by one radiologist and misclassifications were corrected (i.e., discrepancies between DICOM information and actual image content). Table 1 shows the final dataset with all categories selected for the study and the number of images per class. For each of the 30 categories, 100 randomly selected images were set aside for internal validation ($n = 3000$) and the rest of the images was used as the training set ($n = 55{,}219$). To assess the generalizability of the results, we used images from other institutions, acquired with machines from multiple vendors, stored in our PACS ($n = 5324$) as an external validation set. To ensure that these images were labeled correctly, these images were manually labeled by two experienced radiologists because DICOM information could not be automatically processed, was missing, or was in several different languages. These images were not part of the training set and only used to validate the trained network. Table 1 shows the number of images per category for the external validation set.

Neural network training

All images were exported from the PACS as JPEG (Joint Photographic Expert Group) images and anonymized in the process. Using the images in the training set ($n = 58{,}219$), a pretrained MobileNet (Version 1.0) was retrained using oversampling—to account for imbalanced classes—with 22,000 training steps and a learning rate of 0.1. No image augmentation techniques were used. The network was trained on a standard MacBook Pro (Retina, 15–in., Late 2013, 16-GB DDR RAM, 2.3-GHz Quad-Core Intel Core i7).

Statistical analysis

Performance metrics, such as sensitivity, specificity, positive predictive value, and negative predictive value, were calculated using SPSS Version 26.0 [8].

Results

Internal validation

In the internal validation, the overall accuracy of the model in the validation set was 90.3% (95%CI: 89.2–91.3%). Because in this validation set the number of images in each class was equal ($n = 100$), the average sensitivity was the same as the accuracy (90.3%), indicating that, on average, 90.3% of images in each category were correctly classified by the model (see Table 2 for performance metrics for each individual class). As Table 2 shows, the distribution of the sensitivity of the model was rather balanced across categories, ranging between 61.0 and 100.0%. Eighteen out of 30 categories (60.0%) reached a sensitivity of over 90.0%, and 27 out of 30 categories (90.0%) reached a sensitivity of over 80.0%. Only the categories ankle lateral (sensitivity: 79%), lumbar

Table 1 Images per category used for training the network, internal validation, and external validation

Category	Images from own institution	Training	Internal validation	External validation
Abdomen AP	1743	1643	100	218
Abdomen left lateral decubitus	340	240	100	29
Ankle AP	1236	1136	100	75
Ankle lateral	1200	1100	100	94
Cervical spine AP	1209	1109	100	100
Cervical spine lateral	1330	1230	100	150
Chest lateral	7480	7380	100	981
Chest PA/AP	14,217	14,117	100	1114
Elbow AP	1060	960	100	135
Elbow lateral	1136	1036	100	123
Finger AP	664	564	100	30
Finger lateral	793	693	100	28
Foot AP	1234	1134	100	126
Foot oblique	1130	1030	100	106
Hand AP	1683	1583	100	220
Hand oblique	1525	1425	100	195
Hip joint oblique lateral	1409	1309	100	105
Knee AP	2095	1995	100	142
Knee lateral	2045	1945	100	118
Lumbar spine AP	2414	2314	100	166
Lumbar spine lateral	3398	3298	100	233
Panoramic Radiograph	475	375	100	4
Patella axial	842	742	100	10
Pelvis AP	2022	1922	100	113
Shoulder AP	1048	948	100	166
Shoulder outlet	867	767	100	117
Thoracic spine AP	778	678	100	87
Thoracic spine lateral	858	758	100	100
Wrist AP	973	873	100	116
Wrist lateral	1015	915	100	123
Total	58,219	55,219	3000	5324

spine lateral (sensitivity: 77%), and shoulder outlet (sensitivity: 61%) reached a sensitivity below 80.0%.

As for the other performance metrics, the model achieved an average specificity of 99.7%, indicating that, on average, 99.7% of images that were not part of a class were correctly labeled as not belonging to that class. The model achieved an average positive predictive value of 90.8%, indicating that out of all images predicted to belong to a certain class 90.8% of images did actually belong to that class. The average negative predictive value of the model was 99.7%.

External validation

In the external validation, the overall accuracy of the model in the unseen validation set was 94.0% (95%CI: 93.3–94.6%). The average sensitivity of the model was 93.2%, indicating that 93.2% of images in each category were correctly classified by the model (see Table 3 for performance metrics for each individual class). The sensitivity ranged between 75.0 and 100.0%. Twenty-three out of 30 categories (76.7%) reached a sensitivity of over 90.0%, and 29 out of 30 categories (96.7%) reached a sensitivity of over 80.0%. Only the category finger lateral (75%) scored below 80.0%.

As for the other performance metrics, the model achieved an average specificity of 99.8%, indicating that, on average, 99.8% of images that were not part of a class were correctly labeled as not belonging to that class. The model achieved an average positive predictive value of 88.6%, indicating that out of all images predicted to belong to a certain class 88.6% of

Table 2 Performance metrics for the internal validation

Category	Number of images	Sensitivity		Specificity		PPV		NPV	
		Percent	95%CI	Percent	95%CI	Percent	95%CI	Percent	95%CI
Abdomen AP	100	89	82.9–95.1	99.7	99.5–99.9	90.8	85.1–96.5	99.6	99.4–99.8
Abdomen left lateral decubitus	100	100	100.0–100.0	99.9	99.8–100.0	98	95.3–100.0	100	100.0–100.0
Ankle AP	100	80	72.2–87.8	99.4	99.1–99.7	82.5	74.9–90.0	99.3	99.0–99.6
Ankle lateral	100	79	71.0–87.0	100	99.9–100.0	98.8	96.3–100.0	99.3	99.0–99.6
Cervical spine AP	100	96	92.2–99.8	100	99.9–100.0	99	97.0–100.0	99.9	99.7–100.0
Cervical spine lateral	100	97	93.7–100.0	99.8	99.7–100.0	95.1	90.9–99.3	99.9	99.8–100.0
Chest lateral	100	100	100.0–100.0	99.9	99.7–100.0	96.2	92.5–99.8	100	100.0–100.0
Chest PA/AP	100	100	100.0–100.0	99.8	99.6–99.9	93.5	88.8–98.1	100	100.0–100.0
Elbow AP	100	89	82.9–95.1	98.9	98.6–99.3	74.2	66.3–82.0	99.6	99.4–99.8
Elbow lateral	100	96	92.2–99.8	99.6	99.3–99.8	88.1	82.0–94.2	99.9	99.7–100.0
Finger AP	100	82	74.5–89.5	99.4	99.1–99.7	82	74.5–89.5	99.4	99.1–99.7
Finger lateral	100	81	73.3–88.7	99.3	99.1–99.6	81	73.3–88.7	99.3	99.1–99.6
Foot AP	100	92	86.7–97.3	99.6	99.3–99.8	87.6	81.3–93.9	99.7	99.5–99.9
Foot oblique	100	95	90.7–99.3	99.9	99.7–100.0	96	92.1–99.8	99.8	99.7–100.0
Hand AP	100	84	76.8–91.2	99.7	99.5–99.9	91.3	85.5–97.1	99.4	99.2–99.7
Hand oblique	100	89	82.9–95.1	99.8	99.6–100.0	93.7	88.8–98.6	99.6	99.4–99.8
Hip joint oblique lateral	100	91	85.4–96.6	99.9	99.8–100.0	96.8	93.3–100.0	99.7	99.5–99.9
Knee AP	100	98	95.3–100.0	99.9	99.7–100.0	96.1	92.3–99.8	99.9	99.8–100.0
Knee lateral	100	93	88.0–98.0	99.6	99.3–99.8	87.7	81.5–94.0	99.8	99.6–99.9
Lumbar spine AP	100	94	89.3–98.7	99.9	99.8–100.0	97.9	95.1–100.0	99.8	99.6–100.0
Lumbar spine lateral	100	77	68.8–85.2	100	99.9–100.0	98.7	96.2–100.0	99.2	98.9–99.5
Panoramic Radiograph	100	99	97.0–100.0	100	100.0–100.0	100	100.0–100.0	100	99.9–100.0
Patella axial	100	100	100.0–100.0	100	100.0–100.0	100	100.0–100.0	100	100.0–100.0
Pelvis AP	100	95	90.7–99.3	99.9	99.8–100.0	96.9	93.5–100.0	99.8	99.7–100.0
Shoulder AP	100	85	78.0–92.0	98.6	98.2–99.0	68	59.8–76.2	99.5	99.2–99.7
Shoulder outlet	100	61	51.4–70.6	99.4	99.1–99.7	78.2	69.0–87.4	98.7	98.2–99.1
Thoracic spine AP	100	97	93.7–100.3	100	100.0–100.0	100	100.0–100.0	99.9	99.8–100.0
Thoracic spine lateral	100	93	88.0–98.0	99.7	99.5–99.9	92.1	86.8–97.3	99.8	99.6–99.9
Wrist AP	100	86	79.2–92.8	99.7	99.5–99.9	90.5	84.6–96.4	99.5	99.3–99.8
Wrist lateral	100	90	84.1–95.9	98.8	98.4–99.2	72.6	64.7–80.4	99.7	99.4–99.9

images did actually belong to that class. The average negative predictive value of the model was 99.8%.

Discussion

The goal of the present study was to create a neural network for practical applications in the imaging pipeline, e.g., to detect and correct errors in DICOM metadata, to rout radiographs to more specialized networks for abnormality detection, to check that exams are not repeated unnecessarily, to control that the acquired image is the same as the one that was ordered, and to streamline hanging protocols for optimal reporting on images. Our trained model was able to correctly classify the most common types of plain radiographs (e.g., thorax pa, abdomen lateral) and showed good generalizability in the internal (average accuracy: 90.3%) and external validation (average accuracy: 94.0%). However, an overall high accuracy does not necessarily mean that a model will be useful under real-world conditions. One important factor is a comparable level of high performance across all different categories. Combining the results from the internal and external validation set, performance across categories was generally balanced, with only four categories, ankle lateral (79.0%), lumbar spine lateral (77.0%), finger lateral (75.0%), and shoulder outlet (sensitivity: 61.0%) scoring below 80.0%. Taking a closer look at the errors in these categories revealed that the model tended to suggest similar categories and that the correct classification was in

Table 3 Performance metrics for the external validation

Category	Number of Images	Sensitivity		Specificity		PPV		NPV	
		Percent	95%CI	Percent	95%CI	Percent	95%CI	Percent	95%CI
Abdomen AP	218	92.2	88.6–95.8	99.7	99.5–99.8	92.2	88.6–95.8	99.7	99.5–99.8
Abdomen left lateral decubitus	29	86.2	73.7–98.8	99.8	99.7–100.0	75.8	61.1–90.4	99.9	99.9–100.0
Ankle AP	75	88	80.6–95.4	99.8	99.7–100.0	89.2	82.1–96.3	99.8	99.7–99.9
Ankle lateral	94	85.1	77.9–92.3	99.8	99.7–99.9	89.9	83.6–96.2	99.7	99.6–99.9
Cervical spine AP	100	99	97.0–100.0	99.7	99.6–99.9	87.6	81.5–93.7	100	99.9–100.0
Cervical spine lateral	150	99.3	98.0–100.0	99.8	99.6–99.9	92.5	88.5–96.6	100	99.9–100.0
Chest lateral	981	99.1	98.5–99.7	99.4	99.2–99.7	97.6	96.6–98.5	99.8	99.7–99.9
Chest PA/AP	1114	89.8	88.0–91.5	100	99.9–100.0	99.9	99.7–100.1	97.4	96.9–97.8
Elbow AP	135	88.9	83.6–94.2	99.9	99.8–100.0	95.2	91.5–99.0	99.7	99.6–99.9
Elbow lateral	123	93.5	89.1–97.9	99.8	99.7–99.9	91.3	86.3–96.2	99.8	99.7–100.0
Finger AP	30	86.7	74.5–98.8	99.6	99.5–99.8	57.8	43.3–72.2	99.9	99.8–100.0
Finger lateral	28	75	59.0–91.0	99.7	99.6–99.8	56.8	40.8–72.7	99.9	99.8–100.0
Foot AP	126	90.5	85.4–95.6	99.8	99.7–99.9	91.9	87.1–96.7	99.8	99.6–99.9
Foot oblique	106	94.3	89.9–98.7	99.7	99.6–99.9	87.7	81.7–93.7	99.9	99.8–100.0
Hand AP	220	97.7	95.8–99.7	100	99.9–100.0	99.5	98.6–100.4	99.9	99.8–100.0
Hand oblique	195	98.5	96.7–100.0	99.9	99.8–100.0	98	96.0–99.9	99.9	99.9–100.0
Hip joint oblique lateral	105	98.1	95.5–100.0	99.9	99.9–100.0	97.2	94.0–100.3	100	99.9–100.0
Knee AP	142	95.8	92.5–99.1	99.7	99.6–99.9	91.3	86.7–95.8	99.9	99.8–100.0
Knee lateral	118	94.1	89.8–98.3	99.8	99.7–99.9	91.7	86.8–96.6	99.9	99.8–100.0
Lumbar spine AP	166	91	86.6–95.3	99.9	99.8–100.0	97.4	94.9–99.9	99.7	99.6–99.9
Lumbar spine lateral	233	91.8	88.3–95.4	99.9	99.8–100.0	96.8	94.5–99.1	99.6	99.5–99.8
Panoramic Radiograph	4	100	100.0–100.0	100	100.0–100.0	100	100.0–100.0	100	100.0–100.0
Patella axial	10	100	100.0–100.0	99.9	99.8–100.0	58.8	35.4–82.2	100	100.0–100.0
Pelvis AP	113	96.5	93.1–99.9	99.9	99.8–100.0	94.8	90.7–98.8	99.9	99.8–100.0
Shoulder AP	166	97	94.4–99.6	99.5	99.3–99.7	86.1	81.1–91.1	99.9	99.8–100.0
Shoulder outlet	117	90.6	85.3–95.9	99.9	99.8–100.0	93.8	89.4–98.2	99.8	99.7–99.9
Thoracic spine AP	87	94.3	89.4–99.1	99.8	99.7–99.9	90.1	84.0–96.2	99.9	99.8–100.0
Thoracic spine lateral	100	98	95.3–100.0	99.5	99.3–99.7	79.7	72.6–86.8	100	99.9–100.0
Wrist AP	116	94	89.6–98.3	99.8	99.7–100.0	93.2	88.6–97.7	99.9	99.8–100.0
Wrist lateral	123	91.9	87.0–96.7	99.6	99.4–99.7	83.1	76.8–89.4	99.8	99.7–99.9

many cases the model's second prediction (see Fig. 1). This may in part be due to suboptimal positioning in some images, for example, where the patient's pain may have limited the radiographer's ability to achieve perfect positioning. In contrast, highly standardized and unambiguous image categories (e.g., abdomen left lateral decubitus, patella axial, and chest pa/ap) showed perfect classification results with accuracies of up to 100.0%.

To further assess the performance of our model, it is important to compare its performance with other approaches to

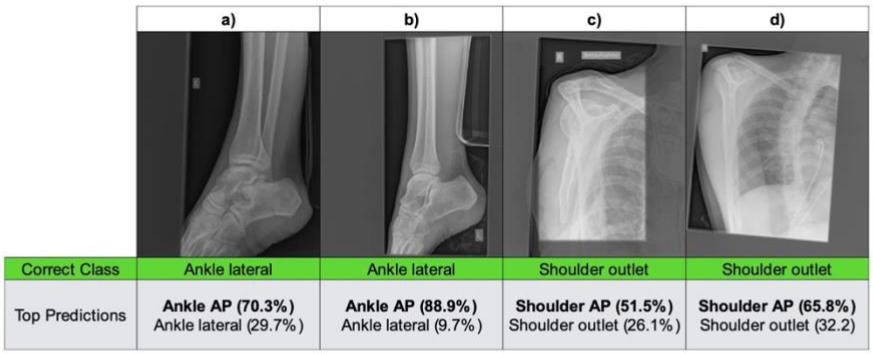

Fig. 1 Examples of four images that were misclassified by the neural network. Images **a** and **b** actually belong to the class ankle lateral but were misclassified as ankle AP by the model. Images **c** and **d** actually belong to the class shoulder outlet but were misclassified as shoulder AP. The corresponding prediction values reflect the probability that the image belongs to a certain class, ranging from 0 to 100%. Higher values reflect a higher probability that an image belongs to a certain class

classify plain radiographs. Using a CNN and Radon transformation, Khatami et al achieved an accuracy of 90.3% for the validation set of the ImageCLEF2009 medical annotation task. This compares favorably with our own accuracy of 90.3% in the internal validation. However, it is difficult to compare performance on different datasets. To allow further assessment of our model, we will make it available so that other institutions are able to test the performance of the model using their own data.

Our study has some limitations: First, even though the 30 categories included in our study accounted for 81.7% (58,219/71274) of all radiographs performed in 1 year, an ideal system should also include the remaining 72 categories.

Second, the overall accuracy was only 90.3% so that every 1 in 10 images would still need some form of human intervention to be correctly classified. There are several reasons for this: (a) Current approaches are relatively "data-hungry," which means they need large amounts of images to achieve a high accuracy. Until new techniques emerge that can produce better results with less data, the only option is for multiple institutions to pool their data for less frequent categories to achieve better performance for rare categories. (b) Performance was generally worse for suboptimal images. As mentioned before, performance of the network will depend on the number of low-quality images in the dataset, as high-quality images with little variation are classified more accurately. Because we did test the network on randomly sampled images from our real PACS, the accuracy of our model may be a more accurate predictor of real-world performance than testing the model on a curated data set with only few low-quality images.

With regard to the accuracy achieved in our study, it is important to note, however, that the errors of the model were not random as the model was particularly prone to mistaking similar categories and the correct option was usually among the top suggestions of the model. Furthermore, it would be feasible to use the probability values generated by the model to flag potentially incorrect predictions because we did find that the probability values for incorrect predictions were significantly lower ($M = 68.2\%$, $SD = 21.0\%$) compared with the probability values for the correct predictions ($M = 95.2\%$, $SD = 10.9\%$) ($t(2708) = 35.7$, $p < .001$, $d = 1.61$).

Taking into consideration the limitations of our model, the following applications for our AI algorithm are feasible: First, the model can be used to classify images and add or correct DICOM metadata. Even though human review is still needed, the workload can be significantly reduced. Considering that very common categories, such as chest pa/ap or chest lateral, were classified with a relatively high accuracy, large parts of a PACS can be corrected with little error. For instance, in our sample, chest imaging accounted for around 30.4% of radiographs performed in 1 year (21,697/71274). With the categories chest pa/ap and chest lateral achieving an accuracy of 100.0% in the internal validation, 30.4% of images in our sample could have been easily labeled using the AI algorithm. Furthermore, being a relatively low-stakes task compared with the detection of abnormalities, it would be relatively safe to deploy the model.

Second, as part of an automated imaging pipeline, the model can be used to route images to more specialized networks for abnormality detection. For instance, the model can first identify a chest image so that it can then be analyzed by a network specialized for detecting anomalies in chest radiographs [9], abdominal radiographs [10], or musculoskeletal radiographs [11–14]. Again, our model did not achieve perfect accuracy for all classes. However, we think that this does not rule out the deployment of the model. One possible solution for this problem would be to use both the average accuracy of a category as well as individual prediction values to decide how to process images. If an image is from a category with high accuracy (e.g., chest pa/ap) and the prediction value for that particular image is high (> 90.0%), it could be sent straight to a secondary network for abnormality detection. If an image is from a category with low accuracy (e.g., shoulder outlet) and the prediction value for that particular image is also low (< 70.0%), it could be flagged for human review.

In summary, we show that it is possible for a single institution to train a neural network to classify the most common categories of plain radiographs, which can then be used to clean up DICOM metadata or as part of an automated imaging pipeline. To encourage independent review and validation as well as to promote the introduction of new tools that may help radiologists and technicians with routine tasks, the final model will be made publicly available on GitHub (https://github.com/healthcAIr/NNCPR).

Funding Open Access funding provided by Projekt DEAL.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Prof. Dr. David Maintz.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was given by all patients before image acquisition at the University Hospital Cologne.

Ethical approval The need for institutional Review board approval was waived since only annoymized retrospective data was used.

Methodology
- Training of a neural network using plain radiographs from the PACS
- retrospective

References

1. Choy G, Khalilzadeh O, Michalski M et al (2018) Current applications and future impact of machine learning in radiology. Radiology 288:318–328. https://doi.org/10.1148/radiol.2018171820
2. Langlotz CP (2019) Will artificial intelligence replace radiologists? Radiology Artificial Intelligence 1:e190058. https://doi.org/10.1148/ryai.2019190058
3. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL (2018) Artificial intelligence in radiology. Nat Rev Cancer 18:500–510. https://doi.org/10.1038/s41568-018-0016-5
4. Dratsch T, Caldeira L, Maintz D, Pinto dos Santos D (2020) Artificial intelligence abstracts from the European Congress of Radiology: analysis of topics and compliance with the STARD for abstracts checklist. Insights Imaging 11. https://doi.org/10.1186/s13244-020-00866-7
5. Harvey H (2018) Why AI will not replace radiologists. https://towardsdatascience.com/why-ai-will-not-replace-radiologists-c7736f2c7d80. Accessed 5 Sept 2020
6. Gueld MO, Kohnen M, Keysers D et al (2002) Quality of DICOM header information for image categorization. In: Siegel EL, Huang HK (eds) Medical imaging 2002: PACS and integrated medical information systems: design and evaluation, pp 280–287
7. Sandström S, Ostensen H, Pettersson H, Akerman K (2003) The WHO manual of diagnostic imaging. World Health Organisation
8. IBM Corp. (2019) IBM SPSS statistics for Macintosh
9. Rajpurkar P, Irvin J, Zhu K et al (2017) CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225
10. Cheng PM, Tejura TK, Tran KN, Whang G (2018) Detection of high-grade small bowel obstruction on conventional radiography with convolutional neural networks. Abdom Radiol 43:1120–1127. https://doi.org/10.1007/s00261-017-1294-1
11. Kim DH, MacKinnon T (2018) Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 73:439–445. https://doi.org/10.1016/j.crad.2017.11.015
12. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S (2018) Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep 8:1727. https://doi.org/10.1038/s41598-018-20132-7
13. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N (2019) Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol 48:239–244. https://doi.org/10.1007/s00256-018-3016-3
14. Üreten K, Erbay H, Maraş HH (2020) Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network. Clin Rheumatol 39:969–974. https://doi.org/10.1007/s10067-019-04487-4

4. Diskussion

4.1. Zusammenfassung der Ergebnisse

Das Ziel dieser Arbeit war es, ein neuronales Netzwerk zur Klassifikation der häufigsten Kategorien von konventionellen Röntgenbildern (z. B.: Thorax ap, Abdomen in Seitenlage) zu entwickeln und anhand von internen und externen Daten zu validieren. Ein solches Netzwerk kann zum Beispiel überprüfen, dass Untersuchungen nicht unnötig mehrfach angefordert werden und dass die erfolgte Untersuchung der angeforderten Untersuchung entspricht. Zusätzlich kann ein solches Netzwerk dabei helfen automatische Aufhängungsprotokolle zu verbessern. Das finale Netzwerk war dazu in der Lage die häufigsten Kategorien von Röntgenbildern zu klassifizieren und zeigte eine gute Generalisierbarkeit im internen (durchschnittliche Genauigkeit: 90.3%) und externen Validierungsset (durchschnittliche Genauigkeit: 94.0%). Hierbei zeigte sich eine ausbalancierte Leistung des Netzwerkes über die verschiedenen Kategorien hin weg, wobei nur vier Kategorien eine Sensitivität von unter 80.0% erreichten: OSG seitlich (79.0%), LWS seitlich (77.0%), Finger seitlich (75.0%) und Schulter seitlich (61.0%). Eine genaue Analyse dieser Kategorien zeigte, dass hierbei das Netzwerk dazu neigte, ähnliche Kategorien miteinander zu verwechseln, zum Beispiel Schulter ap und Schulter seitlich. Ein möglicher Grund hierfür können nicht optimal eingestellte Aufnahmen sein, bei denen es der oder dem MTA nicht möglich war, den Patienten optimal zu positionieren. Dem gegenüber zeigten Kategorien mit mehrheitlich uneindeutigen Bildern (zum Beispiel: Abdomen seitlich, Patella axial und Thorax ap/pa) eine hohe Genauigkeit bei der Klassifikation von nahezu 100.0%.

Vergleicht man die Leistung des finalen Netzwerkes mit anderen Ansätzen, so zeigt sich eine ähnliche Leistung wie zum Beispiel in der Studie von Khatami et al., die mit ihrem neuronalen Netzwerk eine Genauigkeit von 90.3% in dem ImageCLEF2009-Datensatz erreichten[41]. Hierbei ist jedoch zu berücksichtigen, dass die Leistung von neuronalen Netzwerken stark von dem jeweiligen Datensatz abhängt. Trotzdem scheint die Leistung beider Netzwerke mit einer Genauigkeit von jeweils ca. 90.0% vergleichbar zu sein. Um es der wissenschaftlichen Gemeinde zu ermöglichen das in dieser Arbeit entwickelte Netzwerk zu testen und weiterzuentwickeln, wird es online frei verfügbar gemacht.

4.2. Limitierung

Eine der Hauptlimitierungen der aktuellen Studie liegt in der Auswahl der eingeschlossenen Kategorien. Insgesamt beinhaltete der Datensatz zum Training des neuronalen Netzwerkes (alle 71.274 Bilder aus dem Jahr 2017 an unserer Institution) 102 Kategorien an Röntgenbildern. Hieraus wurden die 30 häufigsten Kategorien ausgewählt, die insgesamt 81.7% (58.219 Bilder) der im Jahr 2017 durchgeführten Röntgenbilder an unserer Institution

ausmachten. Die restlichen 72 Kategorien beinhalteten nicht genug Bilder für das Training und die Validierung eines neuronalen Netzwerkes.

Eine weitere Limitierung der aktuellen Studie liegt in der Klassifikationsleistung des neuronalen Netzwerkes: Bei einer Genauigkeit von 90.3% bedeutet dies, dass im Durchschnitt 1 von 10 Bildern nicht korrekt klassifiziert wird. Dies hat verschiedene Gründe: Zum einen benötigen neuronale Netzwerke in ihrer aktuellen Form relativ viele Daten für das Training. Solange nicht neue Techniken zur Verfügung stehen, die genauere Ergebnisse mit weniger Daten erzielen können, bleibt die einzige Option, dass mehrere Institutionen sich zusammenschließen und ihre Daten zusammenlegen, um so eine bessere Leistung für seltenere Kategorien zu erzielen. Zum anderen war die Leistung des neuronalen Netzwerkes schlechter für Röntgenbilder mit geringerer Qualität (zum Beispiel verdrehte Bilder). Insgesamt hängt die Leistung des neuronalen Netzwerkes von der Anzahl an Röntgenbildern mit geringerer Qualität ab, da Bilder mit höherer Qualität und geringer Variation besser klassifiziert werden. Es ist jedoch wichtig hervorzuheben, dass das aktuelle Netzwerk anhand von aus dem PACS unserer Institution zufällig ausgewählten Bildern getestet wurde, so dass die Leistung des Netzwerkes zumindest annähernd realistischen Bedingungen entspricht.

Eine weitere Limitierung des Netzwerkes liegt darin, Bilder ähnlicher Kategorien zu verwechseln. Diesem Problem kann jedoch dadurch begegnet werden, potenziell falsche Kategorisierungen im Vorhinein herauszufiltern. Insgesamt zeigte sich, dass sich die vom Modell generierten Wahrscheinlichkeiten für die einzelnen Kategorien signifikant zwischen korrekten ($M = 68.2\%$, $SD = 21.0\%$) und inkorrekten Klassifizierungen unterschieden, ($M = 95.2\%$, $SD = 10.9\%$) ($t(2708) = 35.7$, $p < .001$, $d = 1.61$). So wäre es zum Beispiel möglich, einen Schwellenwert für Klassifizierungen festzulegen, so dass Röntgenbilder markiert werden, die mit einer Wahrscheinlichkeit unter diesem Schwellenwert klassifiziert werden. Diese Bilder können dann mit menschlicher Hilfe genauer eingeordnet werden. Auch wenn hier vereinzelt noch eine menschliche Überprüfung notwendig ist, kann so trotzdem eine deutliche Arbeitsentlastung erreicht werden, da häufige Kategorien, wie zum Beispiel Thorax pa/ap, mit nahezu 100%-ger Genauigkeit klassifiziert werden und diese einen Großteil der zu klassifizierenden Kategorien ausmachen—zum Beispiel im Falle von Thorax pa/ap ca. 30.4% der an unserer Institution durchgeführten Röntgenbilder in einem Jahr.

4.3. Ausblick

Insgesamt zeigt diese Arbeit, dass neuronale Netzwerke dabei helfen können, häufige Kategorien von Röntgenbildern zu klassifizieren. Neben Röntgenbildern wurden ähnliche Ansätze auch für andere Modalitäten entwickelt. So entwickelten Raffy et al. ein neuronales Netzwerk zur Klassifikation von Körperregionen in CT- und MRT-Aufnahmen[42]. Perspektivisch wäre das Ziel ein neuronales Netzwerk, dass alle möglichen Kategorien von Aufnahmen (z. B.: CT Abdomen, MRT Schulter) in allen radiologischen Modalitäten (Röntgen, CT, MRT,

Mammographie, etc.) erkennen kann. Ein erster Schritt in diese Richtung wurde von Jonske et al. unternommen[43]. Diese entwickelten ein Machine Learning-Modell zur Klassifikation von insgesamt 76 Kategorien für 8 verschiedene radiologische Modalitäten (CT, Angiographie, Röntgen, MRT, PET-CT, PET-MRT, Ultraschall und Mammographie). Während das in dieser Arbeit entwickelte Modell 30 Kategorien von Röntgenbildern klassifiziert, so kann das Modell von Jonske et al. 13 Kategorien von Röntgenbildern klassifizieren. Perspektivisch zeigt sich somit, dass Machine Learning-Modelle zur multimodalen Klassifikation von Bildern in der Radiologie möglich sind und längerfristig die radiologische Arbeit erleichtern können.

4.4. Fazit

Neuronale Netzwerke haben das Potential die radiologische Arbeit grundlegend zu verändern. Während ein Großteil der Forschung in diesem Bereich auf den interpretativen Teil der radiologischen Arbeit fokussiert ist (z. B.: das Erkennen von Anomalitäten und Diagnostizieren von Erkrankungen), sollte diese Arbeit zeigen, dass neuronale Netzwerke vor allem im Bereich der nicht-interpretativen Aufgaben der radiologischen Arbeit einen positiven Beitrag leisten können. Hierzu zählt zum Beispiel die Automatisierung sich wiederholender, einfacher Alltagsaufgaben (z. B.: Messungen, Segmentierungen und das Erkennen von Fremdmaterialien) sowie die Qualitätskontrolle (z. B.: Strahlenschutz oder das Erkennen von suboptimalen Aufnahmen). So kann der Radiologe oder die Radiologin zeitlich und kognitiv entlastet werden, so dass mehr Zeit und kognitive Ressourcen für die Befundung zur Verfügung stehen.

5. Literaturverzeichnis

1 Choy G, Khalilzadeh O, Michalski M, *et al.* Current Applications and Future Impact of Machine Learning in Radiology. *Radiology* 2018; **288**: 318–28.

2 Langlotz CP. Will Artificial Intelligence Replace Radiologists? *Radiol Artif Intell* 2019; **1**: e190058.

3 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018; **18**: 500–10.

4 Dratsch T, Caldeira L, Maintz D, dos Santos DP. Artificial intelligence abstracts from the European Congress of Radiology: analysis of topics and compliance with the STARD for abstracts checklist. *Insights Imaging* 2020; **11**: 59.

5 Harvey H. Why AI will not replace radiologists. 2018. https://towardsdatascience.com/why-ai-will-not-replace-radiologists-c7736f2c7d80 (accessed Sept 5, 2020).

6 Lehmann TM, Güld MO, Deselaers T, *et al.* Automatic categorization of medical images for content-based retrieval and data mining. *Comput Med Imaging Graph* 2005; **29**: 143–55.

7 McCorduck P. Machines Who Think. A K Peters/CRC Press, 2004 DOI:10.1201/9780429258985.

8 TURING AM. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind* 1950; **LIX**: 433–60.

9 Moor J. The Dartmouth College Artificial Intelligence Conference: The next fifty years. *AI Mag* 2006; **27**: 87–91.

10 Gugerty L. Newell and Simon's Logic Theorist: Historical Background and Impact on Cognitive Modeling. *Proc Hum Factors Ergon Soc Annu Meet* 2006; **50**: 880–4.

11 Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958; **65**: 386–408.

12 Haenlein M, Kaplan A. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *Calif Manage Rev* 2019; **61**: 5–14.

13 Zador AM. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat Commun* 2019; **10**: 3770.

14 Muthukrishnan N, Maleki F, Ovens K, Reinhold C, Forghani B, Forghani R. Brief History of Artificial Intelligence. *Neuroimaging Clin N Am* 2020; **30**: 393–9.

15 Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. *Gastrointest Endosc* 2020; **92**: 807–12.

16 Moussa D, Windle G. FROM DEEP BLUE TO DEEP LEARNING : A QUARTER. 2018; **72**: 72–88.

17 Buchanan B. A brief history of artificial intelligence. *AI Mag* 2005; **26**: 53.

18 Markoff J. Computer Wins on 'Jeopardy!': Trivial, It's Not. New York Times. 2011; published online Feb 11. https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html.

19 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017; **60**: 84–90.

20 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**: 436–44.

21 Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci* 1982; **79**: 2554–8.

22 Rajpurkar P, Irvin J, Zhu K, *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017; : 3–9.

23 Cheng PM, Tejura TK, Tran KN, Whang G. Detection of high-grade small bowel obstruction on conventional radiography with convolutional neural networks. *Abdom Radiol* 2018; **43**: 1120–7.

24 Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol* 2018; **73**: 439–45.

25 Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol* 2019; **48**: 239–44.

26 Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. *Sci Rep* 2018; **8**: 1727.

27 Kang G, Liu K, Hou B, Zhang N. 3D multi-view convolutional neural networks for lung nodule classification. *PLoS One* 2017; **12**: e0188290.

28 Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med* 2018; **98**: 8–15.

29 Willemink MJ, Koszek WA, Hardell C, *et al.* Preparing Medical Imaging Data for Machine Learning. *Radiology* 2020; **295**: 4–15.

30 Zou J, Schiebinger L. AI can be sexist and racist — it's time to make it fair. *Nature* 2018; **559**: 324–6.

31 Halpern N, Goldberg Y, Kadouri L, *et al.* Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proc Mach Learn Res* 2018; **81**: 77–91.

32 Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci* 2020; **117**: 12592–4.

33 Gaube S, Suresh H, Raue M, *et al.* Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digit Med* 2021; **4**. DOI:10.1038/s41746-021-00385-9.

34 Schock J, Truhn D, Abrar DB, Merhof D, Conrad S. Automated Analysis of Alignment in Long-Leg Radiographs by Using a Fully Automated Support System Based on Artificial Intelligence. 2021.

35 Kaddioui H, Nahle I, Grimard G. Convolutional Neural Networks for Automatic Risser Stage. 2020.

36 Pan I, Thodberg HH, Halabi SS, Kalpathy-Cramer J, Larson DB. Improving automated pediatric bone age estimation using ensembles of models from the 2017 RSNA machine learning challenge. *Radiol Artif Intell* 2019; **1**: e190053.

37 Elaanba A, Ridouani M, Hassouni L. A Stacked Generalization Chest-X-Ray-Based Framework for Mispositioned Medical Tubes and Catheters Detection. *Biomed Signal Process Control* 2023; **79**: 104111.

38 Patel R, Cantab MA, Thong EHE, *et al.* Automated Identification of Orthopedic Implants on Radiographs Using Deep Learning. 2021; : 1–8.

39 Demircioğlu A, Kim MS, Stein MC, Guberina N, Umutlu L, Nassenstein K. Automatic scan range delimitation in chest ct using deep learning. *Radiol Artif Intell* 2021; **3**. DOI:10.1148/ryai.2021200211.

40 Castiglione J, Somasundaram BSE, Gilligan LA, Trout AT, Brady S. Automated Segmentation of Abdominal Skeletal Muscle on Pediatric CT Scans Using Deep Learning. 2021.

41 Khatami A, Babaie M, Tizhoosh HR, Khosravi A, Nguyen T, Nahavandi S. A sequential search-space shrinking using CNN transfer learning and a Radon projection pool for medical image retrieval. *Expert Syst Appl* 2018; **100**: 224–33.

42 Raffy P, Pambrun J-F, Kumar A, *et al.* Deep Learning Body Region Classification of MRI and CT examinations. 2021; published online April 28. http://arxiv.org/abs/2104.13826.

43 Jonske F, Dederichs M, Kim M-S, *et al.* Deep Learning–driven classification of external DICOM studies for PACS archiving. *Eur Radiol* 2022; published online July 5. DOI:10.1007/s00330-022-08926-w.

6. Anhang

6.1. Abbildungsverzeichnis

7. Vorabveröffentlichungen von Ergebnissen

Basis für diese kumulative Dissertation ist der 2021 in der Fachzeitschrift „European Radiology“ veröffentlichte Artikel „Practical applications of deep learning: classifying the most common categories of plain radiographs in a PACS using a neural network“ mit dem Doktoranden als Erstautor. Im genannten Artikel wurden die Ergebnisse dieser Promotion in Rücksprache mit dem Betreuer vorabveröffentlicht. „European Radiology“ ist eine PubMed gelistete Fachzeitschrift, die Artikel erst nach Peer Review veröffentlicht. Es ist eine offizielle Fachzeitschrift der Europäischen Gesellschaft für Radiologie (ESR). Der Doktorand hatte bei der Erstellung der Publikation den wichtigsten Anteil. So wurde das Manuskript durch den Doktoranden verfasst. Auch Hypothesenstellung, Datensammlung und Datenauswertung erfolgte durch den Doktoranden; hier in Rücksprache mit den Ko-Autor*innen des Papers.