Künstliche Intelligenz in der Radiologie – jenseits der Black-Box

Artificial intelligence in radiology – beyond the black box

Autorinnen/Autoren

Luisa Gallée[1], Hannah Kniesel[2], Timo Ropinski[2], Michael Götz[1, 3]

Institute

1 Division of Experimental Radiology, Department for Diagnostic and Interventional Radiology, University Ulm Medical Centre, Ulm, Germany
2 Visual Computing, University of Ulm, Germany
3 Medical Image Computing, DKFZ, Heidelberg, Germany

Key words

Artificial Intelligence, Explainable AI, Machine Learning, Black Box, Deep Learning, Medical Image Processing

eingereicht 22.12.2022
akzeptiert 22.03.2023
Artikel online veröffentlicht 09.05.2023

Bibliografie

Fortschr Röntgenstr 2023; 195: 797–803
DOI 10.1055/a-2076-6736
ISSN 1438-9029

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Korrespondenzadresse

Prof. Michael Götz
Division for Experimental Radiology, University Ulm Medical Centre, Albert-Einstein-Allee 23, 89081 Ulm, Germany
Tel.: +49/7 31 50 06 11 91
michael.goetz@uni-ulm.de

ZUSAMMENFASSUNG

Hintergrund Die Bedeutung von Künstlicher Intelligenz nimmt in der Radiologie stetig zu. Doch gerade bei neuen und leistungsfähigen Verfahren, vor allem aus dem Bereich des Deep Learnings, ist das Nachvollziehen von Entscheidungen oft nicht mehr möglich. Die resultierenden Modelle erfüllen ihre Funktion, ohne dass die Nutzer die internen Abläufe nachvollziehen können und werden als sogenannte Black-Box eingesetzt. Gerade in sensiblen Bereichen wie der Medizin ist die Erklärbarkeit von Ergebnissen von herausragender Bedeutung, um deren Korrektheit zu verifizieren und Alternativen abwägen zu können. Aus diesem Grund wird aktiv an der Durchleuchtung dieser Black-Boxen gearbeitet.

Methode Dieser Übersichtsartikel stellt unterschiedliche Ansätze für erklärbare Künstliche Intelligenz mit ihren Vor- und Nachteilen vor. Anhand von Beispielen werden die vorgestellten Verfahren veranschaulicht. Die Arbeit soll es dem Leser erlauben, die Grenzen der entsprechenden Erklärungen in der Praxis besser abzuschätzen und die Einbindung solcher Lösungen in neue Forschungsvorhaben stärken.

Ergebnisse und Schlussfolgerung Neben Methoden, Black-Box-Modelle auf Erklärbarkeit zu untersuchen, bieten interpretierbare Modelle eine interessante Alternative. Die Erklärbarkeit ist hier Teil des Verfahrens und das erlernte Modellwissen kann mit Fachwissen überprüft werden.

Kernaussagen:

- Der Einsatz von Künstlicher Intelligenz in der Radiologie bietet viele Möglichkeiten, etwa zur Unterstützung während der Bildaufnahme und -verarbeitung oder zur Diagnosestellung.
- Komplexe Modelle können eine hohe Genauigkeit erreichen, erschweren allerdings die Nachvollziehbarkeit der Datenverarbeitung.
- Wird die Erklärbarkeit bereits bei der Planung des Modells berücksichtigt, können leistungsfähige und zugleich interpretierbare Verfahren entwickelt werden.

Zitierweise

- Gallée L, Kniesel H, Ropinski T et al. Artificial intelligence in radiology – beyond the black box. Fortschr Röntgenstr 2023; 195: 797–803

ABSTRACT

Background Artificial intelligence is playing an increasingly important role in radiology. However, more and more often it is no longer possible to reconstruct decisions, especially in the case of new and powerful methods from the field of deep learning. The resulting models fulfill their function without the users being able to understand the internal processes and are used as so-called black boxes. Especially in sensitive areas such as medicine, the explainability of decisions is of paramount importance in order to verify their correctness and to be able to evaluate alternatives. For this reason, there is active research going on to elucidate these black boxes.

Method This review paper presents different approaches for explainable artificial intelligence with their advantages and disadvantages. Examples are used to illustrate the introduced methods. This study is intended to enable the reader to better assess the limitations of the corresponding explanations when meeting them in practice and strengthen the integration of such solutions in new research projects.

Results and Conclusion Besides methods to analyze black-box models for explainability, interpretable models offer an interesting alternative. Here, explainability is part of the pro-

cess and the learned model knowledge can be verified with expert knowledge.

Key Points:

- The use of artificial intelligence in radiology offers many possibilities to provide safer and more efficient medical care. This includes, but is not limited to support during image acquisition and processing or for diagnosis.
- Complex models can achieve high accuracy, but make it difficult to understand data processing.
- If the explainability is already taken into account during the planning of the model, methods can be developed that are powerful and interpretable at the same time.

Einleitung

Algorithmen der Künstlichen Intelligenz (KI) erlauben die effektive Verarbeitung großer Datenmengen und die Adressierung unterschiedlicher Fragestellungen. Dabei werden in der initialen Trainingsphase bereits bekannte oder bisher verborgene Zusammenhänge in Beispieldaten identifiziert und in einem Modell abgebildet. Mit den so trainierten KI-Modellen werden die gefundenen Korrelationen direkt auf neue Daten angewendet, um diese schnell und einfach zu verarbeiten. Gerade in der Radiologie hat sich dieses Vorgehen aufgrund des hohen Grades der Digitalisierung [1] und der Offenheit für technischen Fortschritt als sehr mächtig erwiesen, um die stetig wachsende Anzahl von Bilddaten [2] auch trotz Fachkräftemangel [3] effektiv verarbeiten zu können.

Das Spektrum der Anwendungen reicht dabei von der effizienten Bildaufnahme über die Optimierung von Arbeitsabläufen hin zu automatischer Diagnoseunterstützung. Beispielsweise ermöglichen KI-Algorithmen eine Reduktion der Messzeit oder der Strahlenbelastung bei gleichbleibender Bildqualität durch eine verbesserte Bildrekonstruktion [4–6]. Eine weitere Anwendung in der täglichen Routine ist die Vorselektion von Bilddaten, um die unnötige Befundung unauffälliger Bilder zu reduzieren. Gerade bei Screening-Programmen wie der Mammografie kann die Arbeitsbelastung signifikant reduziert werden [7–9]. Zudem bieten KI-Verfahren die Chance auf eine schnellere und bessere Diagnose, zum Beispiel durch die automatische Annotation von Organen und Pathologien [10–12] sowie neuen quantitativen und bildbasierten Markern, wie sie aktuell im Bereich Radiomics intensiv erforscht werden [13–15].

Die Fortschritte der KI-Methoden beruhen auf verbesserten Verfahren [16, 17], größeren Datenmengen [18] und gesteigerter Rechenkapazität [19], die die Erzeugung immer komplexerer Modelle ermöglichen. Eine Herausforderung bei dem Einsatz solcher aufwendiger KI-Verfahren ist allerdings die oft schwierige Nachvollziehbarkeit der Entscheidungsprozesse [20, 21]. Insbesondere in der klinischen Routine müssen Entscheidungen, auch solche, die mithilfe von KI-Algorithmen getroffen werden, dringend notwendig nachvollziehbar sein [22]. Gründe dafür sind beispielsweise die Akzeptanz der Patienten und auch die Möglichkeit zur Bewertung der Modellentscheidung.

Das genutzte Wissen wird implizit während des Trainings eines KI-Verfahrens aus den Trainingsdaten gewonnen und auf neue Aufgaben übertragen. Doch dieser Prozess führt zu einigen Unsicherheiten. Wurden alle relevanten Informationen genutzt oder fehlten diese während des Trainings? Sind die gefundenen Korrelationen generalisierbar? Gibt es einen kausalen Zusammenhang für die gefundenen Korrelationen oder sind diese zufällig? Um die Sicherheit eines KI-Systems zu gewähren, muss gezeigt werden, dass das System die zugrunde liegenden Eigenschaften gelernt hat und die Entscheidungen nicht auf irrelevanten Korrelationen zwischen Ein- und Ausgabewerten beruhen, die im Trainingsdatensatz vorkommen.

Durch eine sorgfältige Wahl der Modellarchitektur und des Trainingsalgorithmus eines KI-Verfahrens können Schwachstellen zwar reduziert, aber nicht ausgeschlossen werden. Zusätzliche Informationen helfen, den Einfluss von Störgrößen zu minimieren und die Validierungen der Algorithmen auf externen Datensätzen erlauben die Einschätzung der Generalisierbarkeit und werden zu Recht in datengetriebenen Bereichen wie der Radiomics-Forschung explizit untersucht und gefordert [23, 24]. Doch auch bei sorgfältigem Vorgehen sind Fehler möglich, wie Beispiele aus der Praxis zeigen. So haben Forscher des Mount Sinai Krankenhauses ein KI-Verfahren zur Einschätzung des Pneumonierisikos anhand von Röntgenaufnahmen entwickelt, welches außerhalb dieses Krankenhauses durch signifikant niedrigere Genauigkeiten auffiel. Wie sich herausstellte, nutzte das Verfahren Informationen über die verwendeten Bildgebungsgeräte und erkannte Hochrisikopatienten aufgrund der auf der Intensivstation verwendeten Geräte [25]. Dieses Beispiel zeigt eindrücklich, wie wichtig die Nachvollziehbarkeit eines KI-Systems ist, um derartige Fehlkorrelationen nicht nur durch Zufall, sondern systematisch aufdecken zu können.

Zwischen den einzelnen KI-Verfahren gibt es große Unterschiede nicht nur in der Leistungsfähigkeit, sondern auch bezüglich der Nachvollziehbarkeit der erzeugten Modelle (s. ▶ **Tab. 1**). Sind die Modelle dabei nicht interpretierbar, wird häufig das Bild einer abgeschlossenen, schwarzen Kiste benutzt, die sogenannte Black-Box (s. ▶ **Abb. 1**). Damit werden Modelle beschrieben, deren Funktionsweisen nicht interpretiert werden können, sondern von denen lediglich die Ein- und Ausgabewerte verständlich sind. Um die Funktionsweise einer Black-Box zu verstehen, werden folglich Erklärungsmodelle für das eigentliche Modell benötigt. Im Kontrast dazu stehen interpretierbare Modelle, die entsprechend als White-Box bezeichnet werden. Eine Zwischenstufe beider Extreme ist die Gray-Box. Wir bezeichnen damit Modelle, die gewisse Einblicke über die interne Datenverarbeitung zulassen. Es bleibt zu beachten, dass in der Praxis die Zuordnung zu White-, Gray- oder Black-Box-Verfahren nicht immer eindeutig ist.

White-Box-KI

Optimalerweise ist die gesamte Verarbeitungskette der Daten nachvollziehbar – die entsprechenden Verfahren werden als White-Box-Verfahren bezeichnet. Hier sind vor allem Methoden aus den Bereichen des klassischen maschinellen Lernens und des statistischen Lernens zu nennen, die eine transparente Informationsverarbeitung der Eingabewerte, etwa Patientendaten, Laborwerte oder Bilddaten, hin zu dem Ausgabewert, beispielsweise einer Diagnose, bieten. Ein Beispiel dafür ist die **Lineare Regression**, die eine Linearkombination aus verschiedenen numerischen Merkmalen berechnet. Diese Verfahren werden eingesetzt, um beispielsweise eine Radiomics-Signatur zu bestimmen und die einzelnen Merkmale der Struktur, Form und Textur zu gewichten. Der Einfluss jedes Merkmals wird dabei über ein einzelnes Gewicht bestimmt und kann einfach abgelesen und interpretiert werden [26]. Ähnliches gilt für andere Verfahren wie die **Naive-Bayes-Klassifikation** [27], die durch relative Auftrittswahrscheinlichkeiten von Merkmalen eine Klassenzugehörigkeit schätzt. Durch die Nutzung von Wahrscheinlichkeitsverteilungen ermöglicht der Naive-Bayes-Klassifikator eine einfache Interpretation des Einflusses eines Eingabewerts auf die Modellausgabe.

Transparenz ist jedoch nicht gleichbedeutend mit Interpretierbarkeit. So kann die Interpretierbarkeit auch von White-Box-Verfahren eingeschränkt sein. Deutlich wird dies an **Entscheidungsbäumen** und deren Weiterentwicklung **Random Forests**, die ebenfalls im Bereich Radiomics häufig zum Einsatz kommen [28–31]. Entscheidungsbäume modellieren eine strukturierte Reihe an Konditionen in einer Baumstruktur. Ist der Entscheidungsbaum komplex oder werden Random Forests mit mehreren Bäumen genutzt, sind Entscheidungen zwar transparent und theoretisch nachvollziehbar, in der Praxis aber aufgrund der resultierenden Komplexität nicht mehr [32].

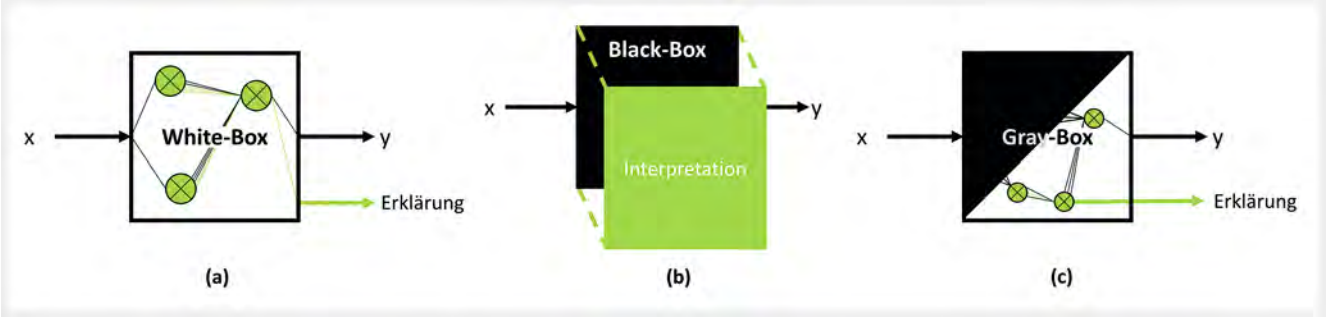
▶ **Tab. 1** Vergleich der unterschiedlichen Leistungsfähigkeiten und Erklärbarkeiten von White-, Black-, und Gray-Box-Verfahren.

	Leistungsfähigkeit	Erklärbarkeit
White-Box	Nur eingeschränkte Komplexität der Modelle	Direkte Interpretation der Modelle gegeben
Black-Box	Komplexe Modelle möglich	Nachträgliche, indirekte Interpretation einzelner Aspekte mittels Erklärungsmodellen
Gray-Box	Komplexe Modelle möglich	Interpretation bezüglich definierter Aspekte durch Modell gegeben, weitere Erklärungen über Black-Box-Verfahren

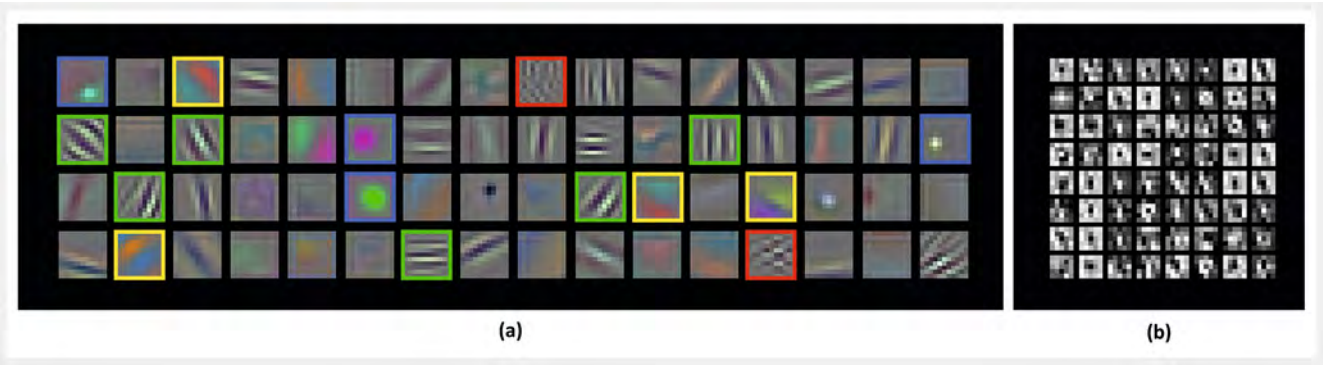
Black-Box-KI

Sind die Entscheidungen eines Verfahrens nicht mehr nachzuvollziehen, etwa aufgrund der Komplexität, spricht man von sogenannten Black-Box-Modellen. Typisch dafür sind Deep-Learning-basierte Verfahren (DL), die oftmals die Leistungen der klassischen Verfahren übertreffen. Sie bilden die Grundlage für die führenden Methoden bei einem breiten Spektrum an komplexen Aufgaben einschließlich der medizinischen Bildanalyse und werden zunehmend in der Radiologie eingesetzt. Deep Learning ist dem Aufbau und der Funktionsweise des Gehirns nachempfunden und nutzt eine engmaschige Verschaltung von Millionen künstlicher Neuronen, die in mehreren Schichten hintereinandergeschaltet sind. Die Verschaltung der Neuronen erlaubt dabei flexible Anpassungen an die jeweilige Aufgabe, wobei die eingegebenen Bilder innerhalb des neuronalen Netzes zu visuellen Merkmalen verarbeitet und damit Segmentierungen erzeugt oder Klassifikationen durchgeführt werden. Die künstlichen Neuronen, in denen das Modellwissen gespeichert ist, sind durch erlernbare Parameter definiert.

Durch die hohe Anzahl an Parametern sind Deep-Learning-Modelle de facto nicht mehr nachvollziehbar [33] und neue Methoden zur Nachvollziehbarkeit des Entscheidungsprozesses sind notwendig. Um die Black-Box des Deep Learnings transparenter zu gestalten, werden deshalb Verfahren entwickelt, welche gezielt die unklaren Funktionalitäten und Verschaltungen der neuronalen Netze zu erklären versuchen. Viele dieser Methoden können auf aktuelle DL-Verfahren aus der allgemeinen Bildverarbeitung angewendet werden. Allerdings variiert die Aussagekraft und der Beitrag zur Interpretierbarkeit der Methoden. Werden die Grenzen dieser Methoden nicht berücksichtigt, besteht die Gefahr einer scheinbaren Erklärbarkeit und dem Ableiten falscher Schlüsse.



▶ **Abb. 1** Schematische Darstellung von (**a**) White-Box-, (**b**) Black-Box- und (**c**) Gray-Box-Verfahren. Der Datenverarbeitungsprozess ist bei White-Box-Verfahren transparent, wohingegen für Black-Box-Verfahren nur Interpretationsmodelle erstellt werden können, die wiederum Fehlerquellen bieten. Als Gray-Box können Verfahren bezeichnet werden, die eine komplexe Informationsverarbeitung mit interpretierbaren Modulen vereint.



▶ **Abb. 2** Visualisierung der Merkmalsfilter eines CNN-Models, das zwischen 100 verschiedenen Tieren differenzieren kann. Die Filter der ersten Schicht (**a**) können noch verständlich beschrieben werden (grüne Box: Linienfilter, blaue Box: Kreisfilter, rote Box: Rauschfilter, gelbe Box: Farbfilter), wohingegen den Filtern in der zweiten, tieferen Schicht (**b**) keine verständliche Funktion mehr zugeordnet werden kann.

Die Basis der meisten bildbasierten DL-Architekturen sind sogenannte Convolutional Neural Networks (CNNs), die mit Filtern Bildmerkmale extrahieren. Die **Visualisierung** dieser Filter (s. ▶ **Abb. 2**) kann Aufschluss über die extrahierten Eigenschaften der Bilddaten geben. Filter in frühen Schichten des Netzwerks extrahieren beispielsweise Linien- oder Kreismuster. Filter aus tieferen Schichten können allerdings nur schwierig interpretiert werden. Die Visualisierung von Filtern hat vor allem dazu beigetragen, die Funktionsweise der CNNs genauer zu verstehen und zu verifizieren. Aufgrund des hohen Abstraktionsgrades der Filtervisualisierung ist diese Technik zur Erklärung der Modellausgabe in einem einzelnen Anwendungsfall allerdings nicht hilfreich.

Ein anderer Ansatz ist die Verwendung von **Optimierung**, um ein Eingabebild zu erzeugen, das gewisse Neuronen maximal aktiviert [34]. Wird ein Neuron stark aktiviert, bedeutet das, dass ein Bildmerkmal, das von diesem Neuron gelernt wurde, in dem Eingabebild vorhanden ist. So konvergiert das Verfahren in Bilder, die Muster abbilden, auf die die gewählten Neuronen trainiert wurden. Als Eingabebild kann entweder zufälliges Rauschen optimiert werden, oder aber es werden Bilder aus dem Trainingsdatensatz gesucht, welche die Aktivierung maximieren. Erstere Methode liefert meist nur abstrakte Bilder, die aber während der Modellentwicklung hilfreich sein können. Letztere liefert leichter interpretierbare Bilder, limitiert dagegen die Spezifität, wenn es nicht eindeutig ersichtlich ist, welches Element in den Eingabebildern tatsächlich zu der hohen Aktivierung der Neuronen geführt hat. Dennoch kann dieser Ansatz in der Praxis in manchen Fällen hilfreich sein.

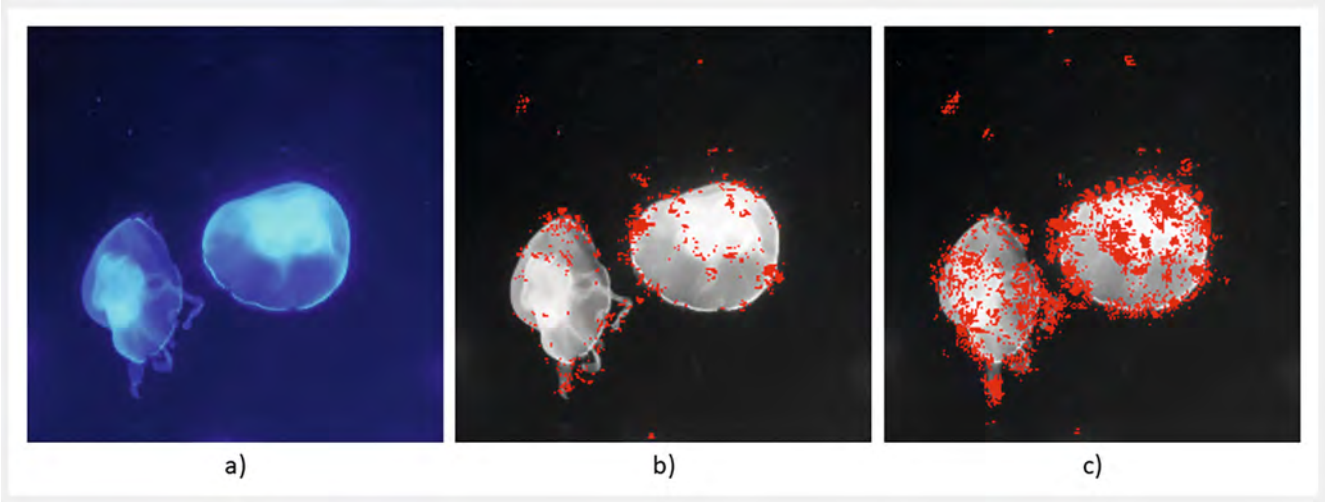
Die **Deconvolution** [35, 36] ist eine approximierte Umkehrung der Convolution eines CNNs. Dabei werden jene Bereiche des Eingabebildes hervorgehoben, die zur Aktivierung einzelner Merkmalsfilter beitragen. Auch hier besteht die Notwendigkeit der Interpretation durch den Menschen, welche Bildmerkmale genau im Bildbereich hervorgehoben werden. Aus diesem Grund und durch die Vielzahl der Filter, die für komplexe Bildanalysen notwendig sind, findet Deconvolution meist nur während der Entwicklung der Modelle zur unterstützenden Analyse Einsatz.

Ungeachtet der inneren Struktur eines Modells untersuchen **maskierungsbasierte Saliency-**Methoden das Modell als echte Black-Box lediglich von außen [37]. Durch gezielte Manipulationen der Eingabedaten und Beobachtung der Veränderung der Ausgabewerte können Zusammenhänge zwischen einzelnen Eingabeparametern und Ergebnissen erstellt werden. Im Kontext der Bildanalyse wird die Manipulation der Eingabe durch Verdecken oder Manipulation einzelner Bildpixel umgesetzt. Im besten Fall erfolgt eine Änderung der Modellausgabe ausschließlich als Reaktion auf das Abdecken relevanter Bereiche. Andernfalls kann auf falsch gelernte Korrelationen geschlossen werden. Des Weiteren können neben räumlicher Bedeutung auch Intensitätseinflüsse überprüft werden. Allerdings ist eine umfassende Überprüfung mit diesem Verfahren aufwendig und selbst bei positivem Ergebnis kann nicht auf Fehlerfreiheit geschlossen werden.

Mit **gradientenbasierten Saliency**-Methoden können Bereiche im Eingabebild hervorgehoben werden, die zur Entscheidung einer gewissen Ausgabe beitragen [38, 39]. Mit diesem Ansatz kann erkannt werden, ob für eine getroffene Entscheidung irrelevante Bildbereiche betrachtet wurden (s. ▶ **Abb. 3**). So wurde bei der Erkennung von COVID-19-Pathologien in Thorax-Röntgenbildern [40] gezeigt, dass der Fokus der gelernten KI auch außerhalb der Lunge und gar des Körpers lag und so Unterschiede in der Patientenpositionierung und der Röntgenprojektion widerspiegelten. Obwohl dieses Feldbeispiel eindrücklich zeigt, dass dieses Verfahren unzureichend generalisierte Deep-Learning-Modelle aufdecken kann, ist bei der Einführung dieser Algorithmen Vorsicht geboten. Denn selbst wenn ein Fokus auf den korrekten Bildbereich festgestellt wird, können inkorrekte Bildmerkmale in diesem Bereich gelernt worden sein und der Einsatz von Saliency-Analysen zu einer Überschätzung des Modells führen.

Ein abstrakterer Ansatz zur Erklärung von Deep-Learning-Modellen ist das **T-CAV**-Verfahren [41]. Es zielt darauf ab, den Einfluss von Konzepten der Eingabebilder zu untersuchen. Ein lineares Klassifizierungsmodell wird trainiert, um unterschiedliche Konzepte basierend auf den Eingabedaten zu lernen. Die Daten können dann ausgehend von ihren Konzepten untersucht werden. Damit kann etwa während der Modellentwicklung ein voreingenommenes Modell frühzeitig erkannt werden. Die Funktionalität von T-CAV hängt allerdings stark von dem trainierten Modell und den resultierenden Konzepten ab.

Die vorgestellten Methoden zeigen, wie unterschiedlich die Erklärungsansätze für Deep-Learning-Netzwerke sind. Grundsätz-



▶ **Abb. 3** Visualisierung der Saliency Heatmaps des CNN-Models für ein Eingabebild (**a**). Rote Pixel der Heatmaps (**b**) und (**c**) zeigen Bildbereiche, die großen Einfluss auf die Netzwerkausgabe haben. Heatmap (**b**) zeigt die Fokussierung des Modells für die korrekte Ausgabe „Qualle", die überwiegend auf dem Körper des Tieres liegt. Diese Fokussierung ist allerdings fast identisch für eine falsche Netzwerkausgabe, wie in (**c**) für die Klasse „Kolibri".

lich können diese einen wichtigen Beitrag zur Erklärung von Black-Box-Modellen liefern, weisen aber immer systematische Einschränkungen auf. Zur Erklärung der komplexen Modelle ist immer eine Reduktion notwendig, die zwingend mit einem Verlust an Informationen einhergeht und somit nur Teilerklärungen liefert. Zusammengefasst spricht für die genannten Methoden die gute Anwendbarkeit auf Black-Box-Modelle. Zu den Einschränkungen gehören die limitierte Aussagekraft und die damit einhergehenden Unsicherheiten.

Gray-Box-KI

Durch Gray-Box-Verfahren können die Vorteile von interpretierbaren White-Box-Verfahren mit der Mächtigkeit von Black-Box-Verfahren kombiniert werden. In diesem jungen Forschungsfeld wird die Erklärbarkeit bereits bei der Entwicklung der KI-Verfahren berücksichtigt, um Erklärungsziele ohne nennenswerten Verlust der Leistungsfähigkeit zu erreichen.

Eine Möglichkeit zur Erklärbarkeit ist die Nutzung exemplarischer Beispiele, sogenannter **Prototypen**. Motiviert von dem menschlichen Vorgehen, Vorhersagen zu treffen, werden Entscheidungen anhand der ähnlichsten Beispiele getroffen, die eine direkte Analyse erlauben. Dabei können entweder ganze Bilder oder einzelne Ausschnitte als Prototypen gelernt werden. Solche Systeme erlauben nicht nur die Klassifikation medizinischer Bilder, sondern zeigen gleichzeitig die ähnlichsten Bilder der Trainingsdatenbank [42, 43]. Die Gültigkeit der Modellschätzung kann damit eingeordnet werden und schafft Vertrauen bei dem verantwortlichen Endnutzer. Gleichzeitig können gefundene Prototypen als Schulungsmaterial dienen.

Invertierbare neuronale Netze weisen eine umkehrbare Architektur auf, sodass Ein- und Ausgabe eines Modells getauscht werden können. Diese Umkehrbarkeit kann genutzt werden, um einzelne Schichten des Netzwerkes zu überprüfen. Durch Manipulierung relevanter Merkmale können kontrafaktische Beispielbilder generiert werden, die Aussagen wie *ohne Merkmal A ist das Ergebnis …* erlauben. Genutzt wird diese Technik bereits in der computerassistierten Chirurgie, um die Unsicherheit bei Durchblutungsschätzungen in der Endoskopie zu bestimmen [44]. Auch wenn invertierbare neuronale Netze die möglichen Netzwerkstrukturen begrenzen, bieten sie eine gute Alternative, um KI-Modelle besser zu verstehen.

Der Vorteil von Gray-Box-Verfahren besteht in der Kombination von Nachvollziehbarkeit bei gleichzeitig hoher Leistungsfähigkeit, was gerade in sensiblen Bereichen wie der Medizin wichtige Eigenschaften sind. Allerdings existieren bisher nur für wenige Anwendungsfälle entsprechende Verfahren. Zudem ist auch bei diesen Verfahren die Erklärbarkeit auf spezifische Elemente begrenzt. Wie bei allen Erklärungsverfahren macht es zum Beispiel einen Unterschied, ob einzelne Fälle betrachtet werden, oder ob eine generelle Aussage getroffen werden soll. Abhängig davon müssen unterschiedliche Erklärungsansätze genutzt werden. Aus diesem Grund wird noch weitere Forschung und Entwicklung auf dem jungen Feld der Gray-Box-Verfahren benötigt, um diese passgenau in vielen Anwendungsgebieten einsetzen zu können. Nur dann können auch die Vorteile dieser Verfahren ausgespielt werden.

Zusammenfassung

Künstliche Intelligenz kann einen wichtigen Beitrag zu einer sicheren und effizienteren Radiologie leisten. Doch für die breite Akzeptanz solcher Systeme in der Ärzteschaft, aber auch bei Patienten, ist die Nachvollziehbarkeit von Entscheidungen eine wichtige Voraussetzung. Nur wenn die genutzten Modelle verständlich sind, können Radiologen und Radiologinnen weiterhin ihrer ärztlichen Sorgfaltspflicht nachkommen und fundierte Diagnosen stellen, Patienten umfassend informieren und Entscheidungen begründet dokumentieren. Nicht nur, aber auch, um eine rechtliche Nachvollziehbarkeit getroffener Maßnahmen zu gewährleisten, stellt die Erklärbarkeit der Modelle eine wichtige

Anforderung für die Anwendbarkeit dar. Gerade leistungsfähige Systeme wie Deep-Learning-basierte Algorithmen sind oft zu komplex, um verständlich zu sein. Die Notwendigkeit, interpretierbare Modelle zu schaffen, wurde bereits erkannt und wird aktuell insbesondere durch Methoden, die im Nachhinein auf fertig trainierten Modellen angewendet werden können, mit verschiedenen Ansätzen adressiert. Die Fortschritte der letzten Jahre haben hier zu beachtlichen Weiterentwicklungen geführt, die unterschiedliche Level an Transparenz bieten und die Beantwortung verschiedener Fragen erlauben, ohne die Komplexität der Modelle einzuschränken. Doch die Analyse von außen limitiert die Aussagekraft über das Black-Box-System und die entsprechenden Verfahren können jeweils nur Erklärungsmodelle der Modelle liefern. Diese sind zwingend Reduktionen der ursprünglichen Modelle und deshalb ebenfalls eine Quelle für Fehler.

Die Nutzung von komplexer, aber interpretierbarer Gray-Box-KI bietet hier eine interessante Alternative. Da die Erklärbarkeit Teil dieser Verfahren ist, entfällt der Zwischenschritt eines Erklärungsmodells. Die erlernten Merkmale können analysiert und mit Fachwissen überprüft werden und bieten eine Entscheidungsgrundlage, auf der der Endnutzer die Vertrauenswürdigkeit der Modellaussage überprüfen kann. Da das Erklärungsverfahren integraler Bestandteil der KI-Lösungen ist, muss dieser Einsatz bereits früh mitbedacht werden, und festgelegt werden, welche Teilaspekte des KI-Modells verstehbar sein sollen. Angepasste Algorithmen sind hier notwendig – auch um die richtige Art der Erklärung zu liefern. Die enge Kooperation zwischen Medizin und Informatik ist folglich von essenzieller Bedeutung, um relevante Fragestellungen zu identifizieren und dafür passgenaue Lösungen zu ermitteln.

Fördermittel

University of Ulm
Baustein (L.SBN.0214)

Interessenkonflikt

Die Autorinnen/Autoren geben an, dass kein Interessenkonflikt besteht.

Literatur

[1] Hricak H. 2016 new horizons lecture: beyond imaging – radiology of tomorrow. Radiology 2018; 286 (3): 764–775

[2] Bundesamt für Strahlenschutz, Hrsg. Röntgendiagnostik: Häufigkeit und Strahlenexposition für die deutsche Bevölkerung. 14. April 2022. Zugegriffen: 24. Oktober 2022. [Online]. Verfügbar unter: https://www.bfs.de/DE/themen/ion/anwendung-medizin/diagnostik/roentgen/haeufigkeit-exposition.html

[3] Attenberger U, Reiser MF. Future Perspectives: Wie beeinflusst künstliche Intelligenz die Entwicklung unseres Berufsfeldes? Radiol 2022; 62 (3): 267–270

[4] Chen Y et al. AI-Based Reconstruction for Fast MRI – A Systematic Review and Meta-Analysis. Proc. IEEE 2022; 110 (2): 224–245. doi:10.1109/JPROC.2022.3141367

[5] Reader AJ, Corda G, Mehranian A et al. Deep Learning for PET Image Reconstruction. IEEE Trans. Radiat. Plasma Med. Sci 2021; 5 (1): 1–25. doi:10.1109/TRPMS.2020.3014786

[6] Willemink MJ, Noël PB. The evolution of image reconstruction for CT – from filtered back projection to artificial intelligence. Eur. Radiol 2019; 29 (5): 2185–2195. doi:10.1007/s00330-018-5810-7

[7] Rodriguez-Ruiz A et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur. Radiol 2019; 29: 4825–4832

[8] McKinney SM et al. International evaluation of an AI system for breast cancer screening. Nature 2020; 577: 89–94. doi:10.1038/s41586-019-1799-6

[9] Kooi T et al. Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal 2017; 35: 303–312. doi:10.1016/j.media.2016.07.007

[10] Gu Z et al. CE-Net: Context Encoder Network for 2D Medical Image Segmentation. IEEE Trans. Med. Imaging 2019; 38 (10): 2281–2292. doi:10.1109/TMI.2019.2903562

[11] Huang H et al. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. in ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain 2020. doi:10.1109/ICASSP40776.2020.9053405

[12] Zhou Z, Siddiquee MMR, Tajbakhsh N et al. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE Trans. Med. Imaging 2020; 39 (6): 1856–1867. doi:10.1109/TMI.2019.2959609

[13] Bera K, Braman N, Gupta A et al. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. Nat. Rev. Clin. Oncol 2022; 19 (2): 132–146

[14] Shin J et al. MRI radiomics model predicts pathologic complete response of rectal cancer following chemoradiotherapy. Radiology 2022; 303 (2): 351–358

[15] Lisson CS et al. Deep Neural Networks and Machine Learning Radiomics Modelling for Prediction of Relapse in Mantle Cell Lymphoma. Cancers 2022; 14 (8): 2008. doi:10.3390/cancers14082008

[16] Guo Y, Liu Y, Oerlemans A et al. Deep learning for visual understanding: A review. Neurocomputing 2016; 187: 27–48. doi:10.1016/j.neucom.2015.09.116

[17] Feng X, Jiang Y, Yang X et al. Computer vision algorithms and hardware implementations: A survey. Integration 2019; 69: 309–320. doi:10.1016/j.vlsi.2019.07.005

[18] Kiryati N, Landau Y. Dataset Growth in Medical Image Analysis Research. J. Imaging 2021; 7 (8): 155. doi:10.3390/jimaging7080155

[19] Thompson NC, Greenewald K, Lee K et al. The Computational Limits of Deep Learning, MIT INITIATIVE ON THE DIGITAL ECONOMY RESEARCH BRIEF Vol. 4, Sep. 2020.

[20] He J, Baxter SL, Xu J et al. The practical implementation of artificial intelligence technologies in medicine. Nat. Med 2019; 25 (1): 30–36. doi:10.1038/s41591-018-0307-0

[21] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med 2019; 25 (1): 44–56. doi:10.1038/s41591-018-0300-7

[22] Tonekaboni S, Joshi S, McCradden MD et al. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. Proceedings of the 4th Machine Learning for Healthcare Conference 2019; 106: 359–380

[23] Götz M, Maier-Hein KH. Optimal Statistical Incorporation of Independent Feature Stability Information into Radiomics Studies. Sci. Rep 2020; 10 (1): 737. doi:10.1038/s41598-020-57739-8

[24] Zwanenburg A et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020; 295 (2): 328–338. doi:10.1148/radiol.2020191145

[25] Zech JR, Badgeley MA, Liu M et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Med 2018; 15 (11): e1002683. doi:10.1371/journal.pmed.1002683

[26] Nasief H et al. A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer. Npj Precis. Oncol 2019; 3 (1): 25. doi:10.1038/s41698-019-0096-z

[27] Wood A, Shpilrain V, Najarian K et al. Private naive bayes classification of personal biomedical data: Application in cancer data analysis. Comput. Biol. Med 2019; 105: 144–150. doi:10.1016/j.compbiomed.2018.11.018

[28] Masoud Rezaeijo S, Ghorvei M, Alaei M. A machine learning method based on lesion segmentation for quantitative analysis of CT radiomics to detect COVID-19. 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS) 2020: 1–5. doi:10.1109/ICSPIS51611.2020.9349605

[29] Chaddad A, Zinn PO, Colen RR. Radiomics texture feature extraction for characterizing GBM phenotypes using GLCM. IEEE 12th International Symposium on Biomedical Imaging (ISBI) 2015: 84–87. doi:10.1109/ISBI.2015.7163822

[30] Haniff NSM, Karim MKBA, Ali NS et al. Magnetic Resonance Imaging Radiomics Analysis for Predicting Hepatocellular Carcinoma. International Congress of Advanced Technology and Engineering (ICOTEN), Taiz, Yemen 2021: 1–5. doi:10.1109/ICOTEN52080.2021.9493533

[31] Wu Q et al. Radiomics analysis of magnetic resonance imaging improves diagnostic performance of lymph node metastasis in patients with cervical cancer. Radiother. Oncol 2019; 138: 141–148. doi:10.1016/j.radonc.2019.04.035

[32] Loyola-Gonzalez O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. IEEE Access 2019; 7: 154096–154113. doi:10.1109/ACCESS.2019.2949286

[33] Leong MC, Prasad DK, Lee YT et al. Semi-CNN Architecture for Effective Spatio-Temporal Learning in Action Recognition. Appl. Sci 2020; 10 (2): 557. doi:10.3390/app10020557

[34] Nguyen A, Dosovitskiy A, Yosinski J et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Adv. Neural Inf. Process. Syst 2016; 29. doi:10.48550/arXiv.1605.09304

[35] Dosovitskiy A, Brox T. Inverting visual representations with convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2016: 4829–4837

[36] Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. Computer Vision – ECCV 2014 D. Fleet, T. Pajdla, B. Schiele, und T. Tuytelaars, Hrsg. Cham: Springer International Publishing, 2014; 8689: 818–833. doi:10.1007/978-3-319-10590-1_53

[37] Park SJ, An KH, Lee M. Saliency map model with adaptive masking based on independent component analysis. Neurocomputing 2002; 49 (1/04): 417–422. doi:10.1016/S0925-2312(02)00637-9

[38] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings, 2014.

[39] Adebayo J, Gilmer J, Muelly M et al. Sanity checks for saliency maps. Adv. Neural Inf. Process. Syst 2018; 31. doi:10.48550/arXiv.1810.03292

[40] DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat. Mach. Intell 2021; 3 (7): 610–619. doi:10.1038/s42256-021-00338-7

[41] Kim B et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). Proceedings of the 35th International Conference on Machine Learning 2018; 80: 2668–2677

[42] Chen C, Li O, Tao D et al. This looks like that: deep learning for interpretable image recognition. Adv. Neural Inf. Process. Syst 2019; 32. doi:10.48550/arXiv.1806.10574

[43] Li O, Liu H, Chen C et al. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. Proceedings of the AAAI Conference on Artificial Intelligence 2018; 32 (1). doi:10.48550/arXiv.1710.04806

[44] Adler TJ et al. Uncertainty-aware performance assessment of optical imaging modalities with invertible neural networks. Int. J. Comput. Assist. Radiol. Surg 2019; 14 (6): 997–1007. doi:10.1007/s11548-019-01939-9