Künstliche Intelligenz in der Medizin und regulatorische Aspekte

Entwicklung und Zertifizierung klinischer KI-Software

Autoren: N. Ahmidi u. L. Mareis

Für eilige Leser

Das derzeitige Interesse an Künstlicher Intelligenz (KI) wird weitgehend durch die beeindruckende Leistung großer Sprachmodelle wie ChatGPT getrieben, was erhebliche mediale Aufmerksamkeit erregt hat. Obwohl bereits zahlreiche KI-Lösungen für verschiedene klinische Anwendungen wie Radiologie, Pathologie, Kaloskopie *und Krebstherapie entwickelt wurden, sind bisher nur wenige in Kliniken implementiert. Das wirft die Frage auf, warum dies der Fall ist. Um diesen Umstand näher zu beleuchten, bietet der vorliegende Artikel einen kurzen Überblick darüber, wie Kl funktioniert, und beschreibt den Prozess der Herstellung und der Zertifizierung eines KI-Systems. Außerdem werden die Herausforderungen skizziert, die bei der Garantie von Zuverlässigkeit und Sicherheit klinischer KI-Systeme auftreten.*

Wie lernt eine Kl ein Konzept in der Medizin?

Das Material für die Entwicklung einer Künstlichen Intelligenz bilden umfangreiche Datenmengen. Um ein Konzept für eine Anwendung im Bereich der Medizin zu erlernen, analysieren KI-Algorithmen große medizinische Datensätze, die Patientenakten, Laborergebnisse und medizinische Bilderenthalten können. Der Algorithmus identifiziert Muster in den Daten und erstellt ein Modell, das diese Muster erkennen und Vorhersagen treffen kann.

Um einen *überwachten* KI-Algorithmus zu trainieren, benötigt man einen großen, vorklassifizierten Datensatz, bei dem jeder Datenpunkt einer bestimmten Kategorie wie Diagnose oder Behandlungsergebnis zugeordnet ist. Der Algorithmus nutzt diesen annotierten Datensatz, um ein Modell zu erstellen, das Muster erkennen und Vorhersagen treffen kann. Nach dem Training wird der Algorithmus mit weiteren, bisher ungesehenen Daten getestet, um zu bestimmen, wie gut und präzise er Ergebnisse vorhersagen kann. Dieser Testprozess ist wichtig, da er hilft, etwaige Problemfelder oder Vorurteile im Algorithmus zu identifizieren. Zur Validierung des KI-Algorithmus soll dessen Leistung mit anderen bestehenden Algorithmen oder menschlichen Experten verglichen werden. Auch werden andere Datensätze verwendet, um die Fähigkeit des Algorithmus zur Verallgemeinerung und zur Vorhersage genauer Ergebnisse dort zu testen. Wenn die Kl andere Algorithmen übertrifft und auf verschiedenen Datensätzen konstant gute Leistungen erbringt, gilt sie als validiert.

Es ist wichtig, zu beachten, dass die Leistung des KI-Algorithmus stark von der Qualität und Vielfalt, der für Training und Prüfung verwendeten Daten abhängt. Daher ist es unerlässlich, sicherzustellen, dass die verwendeten Daten die Bevölkerung repräsentieren und frei von Vorurteilen sind. Zu guter Letzt muss ein menschlicher Experte die Vorhersagen des Algorithmus auf Genauigkeit und Sicherheit überprüfen, bevor er in der klinischen Praxis eingesetzt wird.

Wie entwickeln Forscher KI-Systeme für die Gesundheitsversorgung?

Um KI-Systeme für die Gesundheitsversorgung zu entwickeln, nutzen Forscher eine Vielzahl von Algorithmen, die verschiedene Funktionen ausführen. Zum Beispiel gruppieren Clustering-Algorithmen Patienten mit ähnlichen Merkmalen in Untergruppen, Vorhersagealgorithmen antizipieren das Ergebnis eines diagnostischen Tests, Auto-Encoder- Algorithmen kondensieren große Datensätze in handhabbare Formate, und Reinforcement -Learning-Algorithmen bestimmen die optimale Abfolge von Handlungen zur erfolgreichen Patientenbehandlung. Letztere haben sich

bereits z. B. in Strategiespielen wie Schach gegen Menschen durchgesetzt. KI-Forscherinnen und -forscher modifizieren diese Algorithmen, erfinden neue oder verwenden sie, um bisher unbekannte Muster in großen Datensätzen zu identifizieren, Gewinnstrategien für interaktive Spiele zu entwickeln oder Informationen in ihre grundlegenden Bestandteile zu zerlegen. Diese Arbeit ähnelt der von erfahrenen Ärzten, die seit vielen Jahren praktizieren. Eine Kl erlernt diese Fähigkeiten, indem sie zahlreiche Datenproben analysiert. Ohne Daten ist ihr Einsatz von wenig Nutzen.

Eine Algorithmen-Klasse namens Deep Learning, basierend auf dem Konzept neuronaler Netzwerke, kann u.a. Eingangsdaten unterschiedlicher Formate verarbeiten, wie Zahlenreihen, Text, Bilddaten oder eine Kombination davon. Auch jüngste wissenschaftliche Veröffentlichungen greifen hier auf mehrere Datenquellen zurück, wie z. B. molekulare Informationen, Text, medizinische Signale wie Elektroenzephalogramm (EEG)-Daten, radiologische Bildgebung und sogar multimodale Daten.

Wie wird aus einem in Forschungslaboren entwickelten KI-Algorithmus ein einsatzfähiges medizinisches Produkt?

Der Prozess der Übertragung eines KI-Algorithmus von einem Forschungslabor zu einem in Kliniken verwendeten medizinischen Produkt umfasst mehrere Schritte, einschließlich:

1. Wirksamkeitsnachweis: Der KI-Algorithmus muss sich als wirksam in einem Laborumfeld erweisen. Forscher validieren den Algorithmus typischerweise, indem sie die Leistung mit vorhandenen Methoden oder der Leistung von Fachärzten vergleichen.
2. Vorklinische Tests: Der KI-Algorithmus wird in simulierten klinischen Umgebungen getestet, um sicherzustellen, dass er wie erwartet funktioniert.
3. Klinische Studien: Klinische Studien werden durchgeführt, um die Leistung des KI-Algorithmus in einer klinischen Umgebung zu validieren. Diese Studien erfolgen in der Regel in mehreren Phasen und umfassen eine steigende Anzahl von Patienten.
4. Formelle Genehmigung: Der KI-Algorithmus wird zur regulatorischen Genehmigung bei relevanten Behörden wie der Food and Drug Administration (FDA) in den USA oder der Europäischen Arzneimittel- Agentur eingereicht. Die Regulierungsbehörde überprüft den Algorithmus und genehmigt ihn für den klinischen Einsatz, sofern die erforderlichen Sicherheits- und Wirksamkeitsstandards erfüllt sind.
5. Integration in bestehende Systeme: Sobald der KI-Algorithmus für den klinischen Einsatz zugelassen ist, muss er in bestehende medizinische Systeme integriert werden. Dazu gehört auch sicherzustellen, dass der Algorithmus mit elektronischen Patientenakten und anderen medizinischen Geräten kommunizieren kann.
6. Überwachung nach der Vermarktung: Nachdem der KI-Algorithmus auf den Markt gebracht wurde, muss er auf Sicherheit und Wirksamkeit überwacht werden. Jegliche unerwünschten Ereignisse oder Leistungsprobleme werden dabei an die Regulierungsbehörden gemeldet.

Die ersten beiden Schritte werden typischerweise in wissenschaftlichen Forschungslaboren durchgeführt und in peer-reviewed Fachartikeln dokumentiert. Der Aufwand, der in den Schritten 3 bis 5 gefordert wird, variiert stark je nach Risikoklassifizierung des medizinischen Geräts. Das gilt auch für auf Kl basierende medizinische Software. Im Allgemeinen gibt es dafür vier Klassen:

- Klasse I: Dies sind Niedrig-Risiko-Medizingeräte, bei denen es unwahrscheinlich ist, dass sie Patienten oder Benutzern Schaden zufügen. Beispiele hierfür sind Software-Lösungen zur Verfolgung der Fitness oder der Ernährung. Sie erfordern die geringste regulatorische Kontrolle, und Hersteller müssen sich nur bei der Regulierungsbehörde registrieren las sen.
- Klasse II: Dies sind Mittei-Risiko-Medizingeräte, die ein gesteigertes Risiko für Patienten oder Benutzer darstellen, wie z. B. Software zur Vorhersage des Krankheitsverlaufs oder zur

Überwachung lebenswichtiger Funktionen. Sie verlangen eine höhere regulatorische Kontrolle und müssen einer Vorabüberprüfung durch die Regulierungsbehörde unterzogen werden, um sicherzustellen, dass sie den Sicherheits - und Wirksamkeitsstandards entsprechen.

- Klasse III: Dies sind hochriskante Medizingeräte, die ein signifikantes Risiko für Patienten oder Benutzer darstellen, wie z. B. Software zur Krebsdiagnose oder zur Empfehlung von Behandlungen. Sie bedürfen der strengsten regulatorischen Kontrolle und müssen vor der Markteinführung klinisch getestet sowie von der Regulierungsbehörde zugelassen werden.
- Klasse IV: Das ist eine spezielle Kategorie für Medizingeräte, die zur Unterstützung oder Aufrechterhaltung des menschlichen Lebens oder zur Vermeidung einer lebensbedrohlichen Situation eingesetzt werden. Beispiele hierfür sind Software- Lösungen zur Überwachung von Patienten auf Intensivstationen oder zur Steuerung eines implantierbaren Geräts. Sie erfordern die höchste Ebene regulatorischer Kontrolle, müssen vor der Markteinführung klinisch getestet und von der Regulierungsbehörde zugelassen werden.

Das Klassifizierungssystem für Kl-basierte medizinische Software wurde etabliert. um Patienten und Benutzervor möglichen Schäden zu schützen und die Sicherheit und Wirksamkeit der Software für ihren beabsichtigten Einsatz zu gewährleisten. Der Aufwand, einen KI-Algorithmus auf den Markt zu bringen, variiert je nach Risikoklassifizierung erheblich. Beispielsweise sind für Klasse-1-Geräte minimale Arbeit und finanzielle Investitionen erforderlich, während bereits für Klasse-2-Geräte Kosten von etwa einer halben Million Euro pro Produkt anfallen können. Die Aufwendungen für höhere Risikoklassifizierungen können leicht auf mehrere Millionen Euro steigen.

Wo existieren Problemstellen bei der Gewährleistung des "korrekten Verhaltens" von Kl-basierter medizinischer Software?

Im Folgenden beschränken wir uns auf potenzielle Fehler, die bei der Verwendung von KI-Systemen in klinischen Einstellungen auftreten können. Falls ein Arzt die Informationen eines Patienten, wie Blutwerte oder CT-Bilder, in ein KI-System eingibt. muss die Erwartung sein, dass das System eine genaue Diagnose für den Patienten liefert. Es gibt jedoch verschiedene Schritte, die bei diesem Prozess Probleme erzeugen können. Die Frage lautet daher, welche Vorkehrungen getroffen werden können, um diese potenziellen Probleme sicher zu erkennen und zu beheben.

Einige potenzielle Probleme, die berücksichtigt werden müssen, sind Eingabefehler wie Tippfehler. Zusätzlich können unterschiedliche Einheiten oder Standards, die zur Dokumentation von Blutwerte-Ergebnissen verwendet werden, zu Verwirrungen und Ungenauigkeiten führen. Andere Probleme sind Ärzte ohne Zugang zu bestimmten Geräten wie MRl-Maschinen oder Gen-Tests, was zu einer unvollständigen oder unzureichenden Datenlage führen kann, mit der das KI-System gespeist wird. Zudem können, wie bei jeder anderen Software auch, KI-Systeme während des Entwicklungsprozesses Programmfehler enthalten. Obwohl viele der Bedenken, die mit Kl-basierten medizinischen Software-Systemen verbunden sind, bereits durch bestehende Vorschriften für herkömmliche Software berücksichtigt werden, gibt es weiterführende Problemstellen. Die EU-Verordnung 2017/745 setzt u.a. Anforderungen für den gesamten Software-Entwicklungsprozess fest. um zu gewährleisten. dass er sicher und benutzerfreundlich ist. Dazu gehört die Durchführung von Risikobewertungen, die Implementierung eines Qualitätsmanagementsystems und die Einholung einer Bewertung eines Ethikkomitees. Um eine Zertifizierung zu erhalten, muss ein umfassendes Dokument bei einer unabhängigen, benannten Stelle eingereicht werden, die das Produkt testet. Das Testen, Überprüfen und Validieren der Funktionalität der Software ist entscheidend für die Gewährleistung der Patientensicherheit und zuverlässiger Ergebnisse während des klinischen Einsatzes. Allgemein spielt die EU-Richtlinie eine zentrale Rolle bei der Zertifizierung von klinischer

Software. Gemäß dieser Richtlinie müssen Entwickler ihre gewählte Methode rechtfertigen und spezifische Informationen über den Algorithmus liefern, die unabhängig überprüft werden können. Für Algorithmen, die eine Reihe von Messungen verarbeiten, könnten somit Werte definiert werden, um sicherzustellen, dass die Algorithmen sichere, gültige und robuste Ergebnisse für Eingabewerte aus diesem Bereich liefern.

Jedoch können diese Analysen die Patientensicherheit nicht gänzlich sicherstellen. Patientinnen und Patienten sind äußerst komplex und können nicht vollständig durch eine Reihe von Messungen charakterisiert werden. Daher ist es unmöglich, extreme Punkte zu manipulieren und zu testen, um die Sicherheit für die gesamte Bevölkerung zu bestätigen. Dies wirft die Frage auf, wie sichergestellt werden kann, dass der Begründungsprozess ausreichend ist, um die Patientensicherheit zu gewährleisten. ln der Konsequenz muss daher jeder Teil der Software getestet werden. Die Auswahl der als verlässlich bewerteten Teilkomponenten und Methoden kann dann schrittweise erweitert werden. Auch kann eine explizite Modeliierung der berücksichtigten Daten und Prozesse eine Möglichkeit sein, die Zuverlässigkeit und Verständlichkeit von klinischer Software zu verbessern.

Obwohl KI-Systeme in vielen Anwendungen bemerkenswerte Erfolge erzielt haben, gibt es immer noch erhebliche Herausforderungen bei der Gewährleistung, dass sie sich robust und zuverlässig verhalten. Zum Beispiel kann ein KI-System gut auf einem gegebenen Datensatz, während der Trainings- und Testzeit funktionieren, aber es kann unvorhersehbar reagieren, wenn es mit Daten aus einer neuen Umgebung konfrontiert wird. Um dieses Problem anzugehen, ist es wichtig, geeignete Benchmarking-Tools zu haben, die die Robustheit und Zuverlässigkeit von KI-Systemen bewerten können. "Robuste Kl" bezieht sich auf die Fähigkeit eines KI-Systems, unter verschiedenen Bedingungen und bei verschiedenen Arten von Eingaben, einschließlich unerwarteter oder außerhalb der Trainingsdaten liegender Eingaben, konsistent und genau zu arbeiten. Zum Beispiel muss ein autonomes Fahrzeug in der Lage sein, eine Vielzahl unerwarteter Szenarien wie plötzliche Hindernisse oder widrige Wetterbedingungen zu erkennen und angemessen darauf zu reagieren. Allerdings konzentrieren sich aktuelle Benchmarking-Werkzeuge oft auf eingeschränkte Aspekte der KI-Leistung, wie die Genauigkeit auf einem bestimmten Daten Datensatz, und erfassen möglicherweise nicht das volle Spektrum der Herausforderungen, mit denen ein KI-System in realen Situationen konfrontiert sein kann. Deshalb besteht Bedarf an umfassenderen Benchmarking-Werkzeugen, die die Robustheit und Zuverlässigkeit von KI-Systemen in einem breiteren Spektrum von Szenarien und Kontexten bewerten können.

Der Prozess der Zertifizierung von klinischer Software, insbesondere solcher, die Maschinelles Lernen nutzt unterscheidet sich erheblich von der traditionellen Zertifizierung von Produkten. Im Gegensatz zur Zertifizierung eines Autos, bei der ein standardisierter Test- und Zertifizierungsprozess auf jedes Fahrzeug angewendet wird, erfordert die klinische Software-Prüfung einen individualisierten Ansatz für jeden eingereichten Fall aufgrund der einzigartigen Methoden und Implementierungen, die verwendet werden. Hierzu sind geschulte Software- und Experten für Maschinelles Lernen erforderlich, um die Software zu bewerten und auf spezifische Kriterien zu prüfen. Dieser Prozess ähnelt dem Peer-Review Prozess der Wissenschaft. Um den Zertifizierungsprozess zu vereinfachen, könnte ein standardisierter Ansatz ähnlich dem bei der Zertifizierung von Autos angenommen werden. Hier wird der Testprozess in kleinere standardisierte Teilprozesse für Autohersteller aufgeteilt, was zu einem unkomplizierten Zertifizierungsprozess führt. Die Implementierung eines ähnlichen Ansatzes zur Standardisierung könnte helfen, die Prüfung und Zertifizierung von KI-Software zu erleichtern.

Ausblick

In den nächsten Jahren werden mehr medizinische KI-Produkte auf den Markt gebracht werden, und die wissenschaftliche Forschung wird mit Regulierungsbehörden zusammenarbeiten müssen, um die Sicherheit der Produkte zu garantieren. Nur unter den richtigen Rahmenbedingungen kann diese Synergie einen Mehrwert für die medizinische Versorgung der Patienten bieten.