Künstliche Intelligenz in der Medizin: Wo stehen wir heute, und was liegt vor uns?

Augusto Garcia-Agundez und Carsten Eickhoff

Abstract

Künstliche Intelligenz (KI) und maschinelles Lernen (ML) tragen seit Jahrzehnten zur medizinischen Forschung bei und haben sich von rein theoretischer Forschung zu praktischen, wirkungsvollen Werkzeugen entwickelt, die in verschiedenen klinischen Entwicklungsprozessen eingesetzt werden. Dieser Artikel gibt einen Überblick über den aktuellen Stand der klinischen KI und konzentriert sich insbesondere auf sogenannte Natural Language Processing (NLP)-Technologien, die in klinischen Anwendungsfällen eingesetzt werden. Wir skizzieren jede Phase des Lebenszyklus eines klinischen KI-Modells, von der Datenerfassung und -vorverarbeitung über das Training und die Evaluierung bis hin zum praktischen Einsatz und schlussendlich Ersatz. Wir erörtern auch Fähigkeiten, Transparenz, Erklärbarkeit und spezifische Anwendungen, die heute eingesetzt werden. Mit dieser Analyse wollen wir Kliniker über realistische Erwartungen und praktische Überlegungen informieren und anleiten, wenn sie den Einsatz von KI zur Lösung ihrer Forschungsfragen und klinischen Bedürfnisse in Erwägung ziehen.

1 Einleitung

Künstliche Intelligenz (KI) und maschinelles Lernen (ML) tragen in verschiedenen Formen seit Jahrzehnten zur medizinischen Forschung bei und haben sich von rein theoretischer Forschung zu praktischen, wirkungsvollen Werkzeugen entwickelt, die derzeit in klinischen Entwicklungsprozessen eingesetzt werden [12]. Im Zuge der Weiterentwicklung von KI-Methoden sind sie in der Lage, komplexe Muster in Daten zu erfassen und zu verstehen, wobei Fortschritte wie neuronale Netze und Deep Learning [19] (Abbildung 1) sowie die Verarbeitung weiterer Datenquellen wie Bilder [32] oder Text [4] zum Tragen kommen. Klinische Implementierungen haben sich mit ihnen verbessert, von relativ einfachen Expert Decision Trees bis hin zu modernen Decision Support Systems, die in der Lage sind, Fragen aus medizinischen Untersuchungen zu beantworten [28], klinische Bildgebung zu befunden [26] oder kausale Schlussfolgerungen zu ziehen [8, 18, 3].

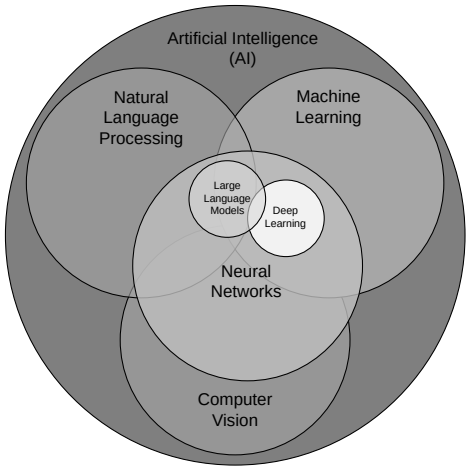

Figure 1: Hierarchie der AI-Architekturen und Begriffe

In den letzten Jahren gab es eine bemerkenswerte Beschleunigung in diesem Bereich durch das Aufkommen von auf großen Sprachmodellen(LLM) basierenden Frage-Antwort-Systemen wie GPT [2]. Diese Modelle, die von den jüngsten Fortschritten im Natural Language Processing (NLP) [35, 6] und dem Instruction Tuning [39] angetrieben werden, machen sich die Tatsache zunutze, dass LLMs schnell große Textmengen als Eingabeabfrage verarbeiten und aus dem Textkontext lernen können, um Ergebnisse wie effiziente Zusammenfassungen und die Beantwortung von Fragen zu liefern.

Trotz dieser jüngsten Fortschritte, und in der Tat zum großen Teil aufgrund dieser Fortschritte, gibt es erhebliche Missverständnisse darüber, was diese Modelle leisten können und was nicht. Phänomene wie Halluzinationen [21], Schmeichelei [33] und andere Formen des Fehlverhaltens von Modellen [11] sind dokumentiert worden und haben bei Angehörigen der Gesundheitsberufe Besorgnis ausgelöst. Selbst wenn sie erwartungsgemäß funktionieren, sind Fragen zu Transparenz, Erklärbarkeit und der so wahrgenommenen "Blackbox"-Natur hochentwickelter ML-Modelle häufige Diskussionsthemen [7].

Um diese Bedenken in einen größeren Zusammenhang zu stellen und zu erläutern, was KI leisten kann und was nicht, gibt dieser Artikel einen Überblick über den aktuellen Stand der klinischen KI und konzentriert sich dabei auf NLP-Technologien, die in klinischen Anwendungsfällen eingesetzt werden. Wir beschreiben den Lebenszyklus eines klinischen KI-Modells, geben einen Überblick über den aktuellen Stand der KI in der Medizin, auch jenseits der Thoraxchirurgie, und erörtern Fähigkeiten, Transparenz, Erklärbarkeit und spezifische Anwendungen, die bereits eingesetzt werden. Mit dieser Analyse wollen wir Mediziner über realistische Erwartungen und praktische Erwägungen informieren und anleiten, wenn sie den Einsatz von KI zur Lösung ihrer Forschungsfragen und klinischen Bedürfnisse in Erwägung ziehen.

2 Lebenszyklus eines KI-Modells

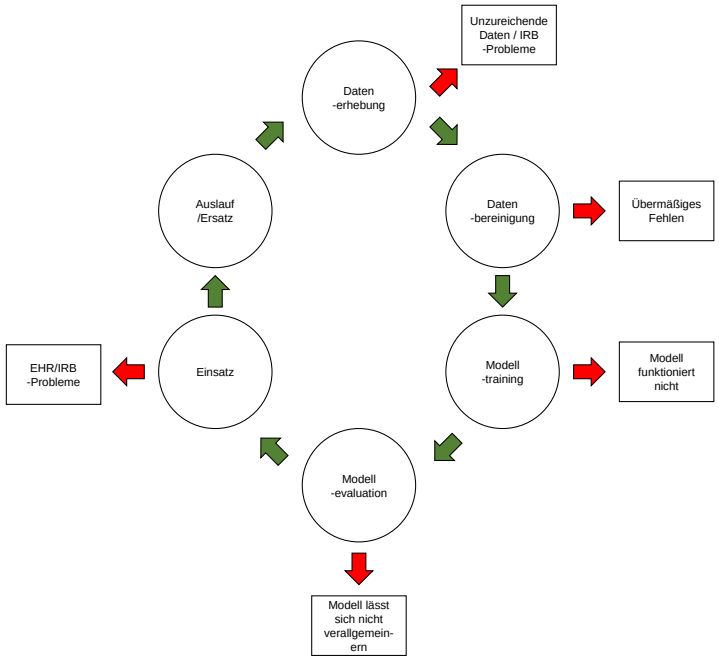

Figure 2: Lebenszyklus eines KI-Modells

Der Lebenszyklus eines klinischen KI-Modells ist ein komplexer mehrstufiger Prozess, der verschiedene Phasen umfasst, von der Datenerfassung bis zum Auslauf oder Ersatz. Im Folgenden werden die einzelnen Phasen des Lebenszyklus beschrieben und die Prozesse untersucht, die die Modelle durchlaufen, um sicherzustellen, dass sie effektiv und zuverlässig sind und allen regulatorischen Anforderungen genügen.

Datenerhebung. Die Entwicklung eines klinischen KI-Modells beginnt mit der Datenerfassung, einer Phase, in der die relevanten Daten zum Trainieren und Evaluieren des Modells gesammelt oder erstellt werden. Im Gesundheitswesen stammen diese Daten in der Regel aus großen elektronischen Gesundheitsakten, auf Englisch Electronic Health Records (EHRs). Klassischerweise beschränkte sich die Datenerfassung auf strukturierte Datenfelder (z. B. Alter, Geschlecht, Raucherstatus, Medikamente, Dosierungen, Laborergebnisse, Vitalparameter, Genomik usw.). Modernere KI-Modelle beziehen jedoch auch andere Datenquellen mit ein, die sehr viel umfangreicher sind, aber eine spezielle Verarbeitung erfordern, wie z. B. Texte oder Bilder. Die Datenerfassung ist ein entscheidender Schritt, da die Funktionalität des Modells - vorausgesetzt, es kann ein solches Modell erstellt werden - vollständig von der Qualität und Quantität der Daten abhängt.

Die Anforderungen an “Datenqualität” und “Datenmenge” sind zu Projektbeginn schwer zu bestimmen. Im Allgemeinen können Forscher lediglich Näherungswerte für diese beiden Faktoren festlegen, indem sie einige der folgenden Fragen beantworten. Würde der Stichprobenumfang für eine konventionellere Analysemethode wie Regression ausreichen? Stelle ich alle Daten zur Verfügung, von denen ich aufgrund meines Fachwissens zumindest annehme, dass sie für das vorliegende Problem relevant sind? Repräsentieren meine Daten die potenzielle Vielfalt von Inputs und Outputs/Ergebnissen adäquat?

Häufige Gründe für ein Scheitern in dieser Phase sind unzureichende Daten oder Hindernisse, auf Daten zuzugreifen und sie teilen oder nutzen.

Datenbereinigung. Nach der Erfassung müssen die Daten vorverarbeitet werden, um ihre Eignung für das Modelltraining zu gewährleisten. In dieser Phase geht es vor allem darum, Ungenauigkeiten und Inkonsistenzen zu beseitigen und fehlende Daten zu korrigieren, um die Ausbreitung von Fehlern zu verhindern. Eine sorgfältige Vorverarbeitung ist unerlässlich, um Verzerrungen zu beseitigen, die die Objektivität und Genauigkeit des Modells beeinträchtigen könnten.

Normalisierung, Interpolation und Extrapolation oder Dimensionalitätsreduktion sind gängige Techniken bei numerischen Daten, um sicherzustellen, dass die Daten so sauber wie möglich sind, bevor das Modell trainiert wird. Ein besonders wichtiger Schritt in dieser Phase ist das Resampling des Klassengleichgewichts für Klassifizierungsaufgaben. Im medizinischen Bereich ist die Verteilung von Fällen und Kontrollen oft stark auf der Seite der Kontrollen verzerrt, speziell wenn seltene Krankheiten oder Ereignisse betrachtet werden sollen. So ist beispielsweise die Wahrscheinlichkeit, dass ein Patient eine postoperative Sepsis entwickelt [15] oder nach der Entlassung bald wieder ins Krankenhaus eingeliefert wird [4], relativ gering . Dies stellt ein Problem für Klassifikatoren dar, die weitgehend auf maximale Genauigkeit trainiert sind, da sie nicht dafür bestraft werden, dass sie die Mehrheitsklasse (Kontrollen) für alle Eingaben vorhersagen, was kein wünschenswertes Verhalten ist [10]. Dieses Problem kann auf verschiedene Weise entschärft werden. Erstens können die Studienkohorten so aufbereitet werden, dass das Ungleichgewicht der Klassen gemildert wird. Gängige Techniken sind die Reduzierung der Mehrheitsklasse, die Wiederholung von Instanzen der Minderheitsklasse oder die Generierung synthetischer Beispiele der Minderheitsklasse unter Verrauschung vorhandener Beispiele [22]. In einigen Fällen kann sich der Klassenausgleich jedoch nachteilig auf die Gesamtleistung auswirken [34]. Darüber hinaus sollten die Entwickler in Erwägung ziehen, das Modell anhand von Metriken zu bewerten, die für unbalancierte Lernumgebungen repräsentativer sind, wie z. B. die Fläche unter der Precision-Recall-Kurve (AUPRC) oder Macro-F1.

Häufige Gründe für ein Scheitern in dieser Phase sind Hindernisse in der Datenbereinigung (z. B. wegen übermäßiger fehlender Daten).

Modelltraining & Optimierung. Sobald die Daten vorverarbeitet sind, wird das KI-Modell trainiert. Dies kann je nach Art des zu entwickelnden Modells unterschiedliche Formen annehmen. Bei herkömmlichen ML-Aufgaben werden oft

mehrere Modelle parallel trainiert und miteinander verglichen (Decision Trees, Logistic Regression, usw., siehe Abbildung 3). Werden komplexe Muster in den Daten erwartet, können verschiedene Modelle aus der Familie des Deep Learning definiert wird [19], verwendet werden, um Abstraktionen der Eingabedaten zu erstellen (besonders nützlich im Falle von Bild- oder Texteingaben). Dazu gehören Modelle, die nach Mustern suchen, wie z. B. Convolutional Neural Networks (CNNs), Recursive Neural Networks (RNNs) oder Long Short-Term Memory Modelle (LSTMs). In den letzten Jahren hat das Deep Learning durch die Entwicklung der Transformer-Architektur einen beträchtlichen Aufschwung erfahren [35]. Transformers sind konzeptionell einfach, aber in der Lage, sehr komplexe und unterschiedliche Muster in Daten, einschließlich Sprache, in einem Umfang zu verstehen, der bisher nicht möglich war (Abbildung 4). Je größer und ausgefeilter die Architektur wird, desto zeitaufwändiger und rechenintensiver wird das Training von Modellen für konkrete Aufgaben. Aus diesem Grund wird in vielen Fällen das Training "von Grund auf" durch Finetuning ersetzt, ein Ansaty, bei dem ein Modell, das zuvor für einen allgemeinen Zweck vortrainiert wurde (*pretraining*), dann für einen kurzen Zeitraum für eine sehr spezifische Aufgabe weiter trainiert (*fine-tuning*) wird [16]. So wurde beispielsweise aus dem allgemeinen Sprachmodell BERT [9], das mit allgemeinem Text trainiert wurde, ein Modell abgeleitet, das mit Artikeln aus der Pubmed Datenbank feinabgestimmt wurde und BioBERT heißt [20]. Anschließend wurde BioBERT mit klinischem Freitext aus dem MIMIC-III-Datensatz weiter verfeinert, was zum Modell ClinicalBERT [4, 14] führte.

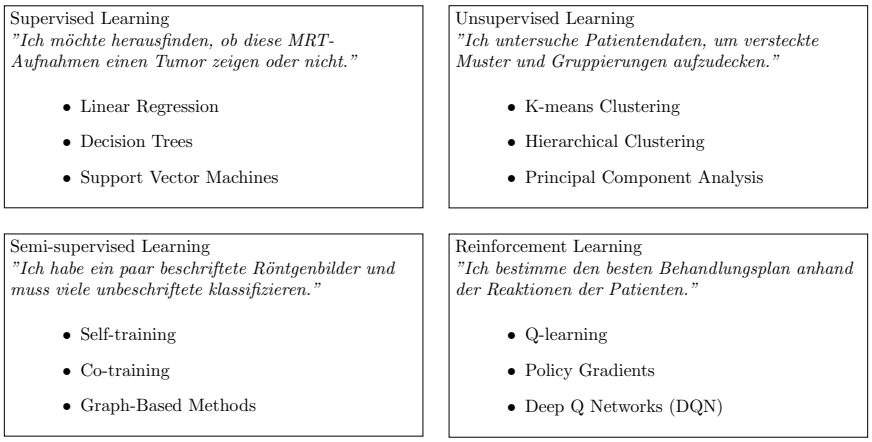

Figure 3: Überblick über Methoden des maschinellen Lernens

Häufige Gründe für ein Scheitern in dieser Phase sind eine unzureichende Modellleistung im weitesten Sinne. So kann ein Modell zwar insgesamt eine gute Leistung erbringen, aber dazu neigen, sich zu stark and die Trainingsdaten anzupassen oder eine übermäßige Anzahl von Falschwarnungen zu produzieren, was zu einer Alarmmüdigkeit (*Alarm Fatigue*) beim Nutzer führen kann.

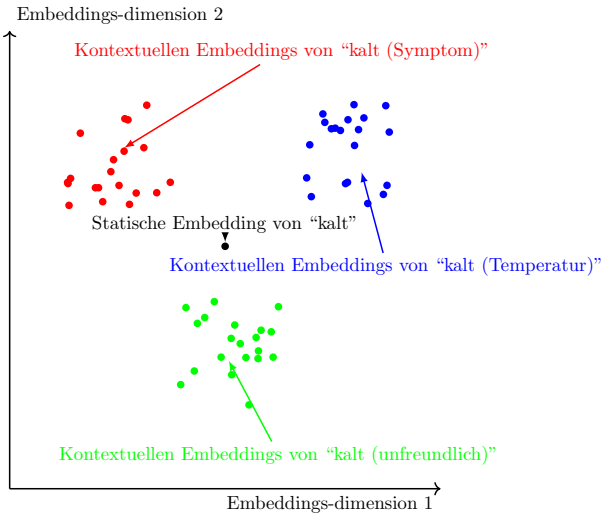

Figure 4: Das kontextuelle Verständnis in Transformers ermöglicht es den Modellen, die Bedeutung eines bestimmten Wortes in verschiedenen Kontexten zu unterscheiden [31]

Modellbewertung Nach dem Training werden die Modelle mit quantitativen Metriken wie Akkuranz, Spezifität und Präzision auf einem Teil der verfügbaren Daten bewertet. In der Regel handelt es sich bei diesen sogenannten Testdaten um einen kleinen Prozentsatz ($10\% - 20\%$) der Gesamtdaten, der als *Test set* definiert ist. Da Modelle viele Stellschrauben haben, die sich auf ihre Funktionsweise auswirken (allgemein "Hyperparameter" genannt), wird ein anderer Teil der Daten, der eine vergleichbare Größe wie der Testsatz hat, verwendet, um verschiedene Parameterwerte zu bewerten, bevor der bisher ungesehene Testsatz verwendet wird. Diese Datenmenge wird gewöhnlich als *Validation set* bezeichnet. Dieser Prozess kann in einem einzigen Datensatz mehrmals wiederholt werden, wobei verschiedene Datenpunkte verschiedenen Datensätzen zugewiesen werden, was als *cross-validation* bezeichnet wird. Diese Bewertung sollte jedoch in Verbindung mit qualitativen Bewertungen durch klinische Experten durchgeführt werden. Dadurch wird sichergestellt, dass die Entscheidungsfindung des Modells mit dem Expertenwissen über das erwartete Verhalten übereinstimmt und das Modell im vorgesehenen klinischen Kontext korrekt funktioniert. Es ist auch wichtig zu bedenken, dass die Leistung eines Modells nicht für alle Patienten gleich sein muss. Beispielsweise könnte ein Klassifikator bei einer bestimmten Klasse deutlich schlechter abschneiden oder bestehende Verzerrungen in den Daten replizieren. In dieser Phase ist es auch wichtig sicherzustellen, dass das Modell auf bisher ungesehene Datenpunkte verallgemeinert

werden kann, indem nach Möglichkeit externe Validierungen in anderen klinischen Zentren durchgeführt werden.

Ein häufiger Grund für das Scheitern in dieser Phase ist, dass das Modell auf ungesehenen Daten nicht gut abschneidet (es versagt bei der Verallgemeinerung). Dies kann daran liegen, dass die Trainingsmethode das Modell dazu veranlasst hat, sich die Daten des Trainingssatzes einzuprägen, anstatt aus ihnen Muster zu lernen (*overfitting*), oder dass die Daten nicht repräsentativ für die allgemeine Kohorte sind.

Einsatz & Regulierung. Erfolgreich evaluierte Modelle werden in den klinischen Arbeitsablauf integriert. In dieser Phase ist die Einhaltung medizinischer Vorschriften und Datenschutzgesetze unerlässlich und erfordert möglicherweise eine Zertifizierung als Medizinprodukt (z.B. CE oder FDA). Der Einsatz umfasst auch die Überwachung und Wartung, um eine gleichbleibende Qualität zu gewährleisten und auftretende technische oder klinische Probleme zu erkennen. Die Schlüsselkomponente dieser Phase ist die Entwicklung einer EHR-schnittstelle (API), die es dem neu trainierten Modell ermöglicht, direkt mit den Versorgungssystemen zu interagieren und die relevanten Daten daraus zu ziehen, im Gegensatz zur anfänglichen Datenerfassung, die manuell erfolgt und in der Regel viel mehr Datenpunkte umfasst, die während der Auswertung als nicht relevant identifiziert werden. Obwohl herkömmliche ML-Modelle stark reguliert sind, wird die Regulierung neuartiger LLMs noch diskutiert. Einige Autoren argumentieren zum Beispiel, dass diese Modelle wie jede andere Interventionsmethode von den relevanten benannten Stellen bewertet werden sollten [5]. Andere Autoren weisen darauf hin, dass sich ihre Regulierung auf Sicherheits- und Gerechtigkeitsprinzipien konzentrieren sollte [24]. Doch wie kann ein Foundation-Modell, das gleichzeitig vielen Zwecken dient, von denen wir uns einige noch gar nicht vorstellen können, mit einem der beiden Schwerpunkte reguliert werden?

Auslaufphase/Ersatz. Da klinische Informationssysteme immer mehr Datenpunkte sammeln und sich die KI-Methoden weiterentwickeln, kann jedes Modell im Laufe der Zeit veraltet sein, selbst wenn sich klinische Entwicklungsprozesse auf sie stützen. Um die maximale Funktionalität aufrechtzuerhalten, ist es wichtig, dass die Ausmusterung und Ersetzung innerhalb eines bestimmten Zeitintervalls erfolgt. Je nachdem, wie kritisch ein Modell für den klinischen Prozess ist, kann dieser Schritt zunehmend komplexer werden, da die Unterbrechung des Betriebs so gering wie möglich gehalten werden sollte.

3 Was kann KI für mich tun?

Sie kann relevante Daten abrufen: LLMs können umfangreiche Patientendaten, die in elektronischen Patientenakten enthalten sind, aggregieren und zusammenfassen, was besonders in solchen Bereichen von Bedeutung ist, in denen Kliniker viel Zeit mit der Durchsicht von Krankenakten verbringen. So können Chirurgen schnell die komplette Krankengeschichte eines Patienten erfassen, einschließlich früherer bildgebender Untersuchungen, Laborergebnisse und chirurgischer Berichte, ohne große Datenmengen manuell durchsuchen zu müssen. Diese Fähigkeit spart nicht nur Zeit, sondern stellt auch sicher, dass keine wichtigen Details übersehen werden, was zu einer fundierteren Entscheidungsfindung und besseren Patientenversorgung führt.

Sie kann die Diagnose unterstützen: ML-Klassifikatoren unterschiedlicher Art sind hervorragende Werkzeuge zur Unterstützung der Differenzialdiagnose. Diese Systeme verwenden beispielsweise Deep-Learning-Algorithmen, die auf großen Datensätzen trainiert wurden, um gutartige von bösartigen Läsionen mit einer Präzision zu unterscheiden, die manuelle Interpretationen übertreffen kann. Auf diese Weise können sie ein früheres Eingreifen ermöglichen, das die Ergebnisse für den Patienten erheblich verbessern kann.

Sie kann prognostizieren: KI-Modelle können eine Vielzahl von Variablen aus den Krankenakten eines Patienten verarbeiten, um therapeutische Ergebnisse vorherzusagen[25]. Indem sie alle verfügbaren EHR-Daten einbeziehen, können sie individuelle Risikobewertungen erstellen. Diese Informationen sind nicht nur potenziell wertvoll für die Entscheidungsunterstützung, sondern auch für die Bereitstellung personalisierter Informationen über Vorteile und Risiken einer Operation für die Patienten, wodurch eine genauere Aufklärung gewährleistet werden kann.

Sie kann klinische Dokumentation beschleunigen: LLMs können die Zeit, die Ärzte mit der Dokumentation verbringen, erheblich reduzieren, indem sie automatisch Berichte, Entlassberichte und andere klinische Dokumente aus diktierten oder getippten Eingaben in Kombination mit klinischen Daten erstellen. Derzeit werden in Pilotprojekten KI-Schreiber für die Dokumentation von Hausarztbesuchen getestet. Diese Technologien können auch bei Abrechnungszwecken helfen und sicherstellen, dass Verfahren und Eingriffe genau erfasst werden, was die Betriebskosten senkt.

Sie kann sich rechtfertigen: Entgegen der landläufigen Meinung sind KI Modelle keine "Black Boxes", die sich nicht erklären lassen. Während weniger komplexe ML-Methoden ihre Funktionsweise leicht mithilfe von Techniken wie Variable Importance Rankings (z. B. [30]) oder Shapley Added Explanations [23] erklären können, stehen für größere Modelle ähnliche Methoden wie die Attention Visualization [36] und Integrated Gradients [1] zur Verfügung. Jüngste Forschungen legen nahe, dass LLMs selbst hervorragende Erklärungswerkzeuge sind [17]. Es stimmt zwar, dass es nicht möglich ist, den Beitrag jedes einzelnen der mehreren Milliarden Parameter des Modells zur Beziehung zwischen Input und Output zu verstehen, doch sollte dies nicht als generelle Unmöglichkeit missverstanden werden, die Funktionsweise der Modelle zu verstehen.

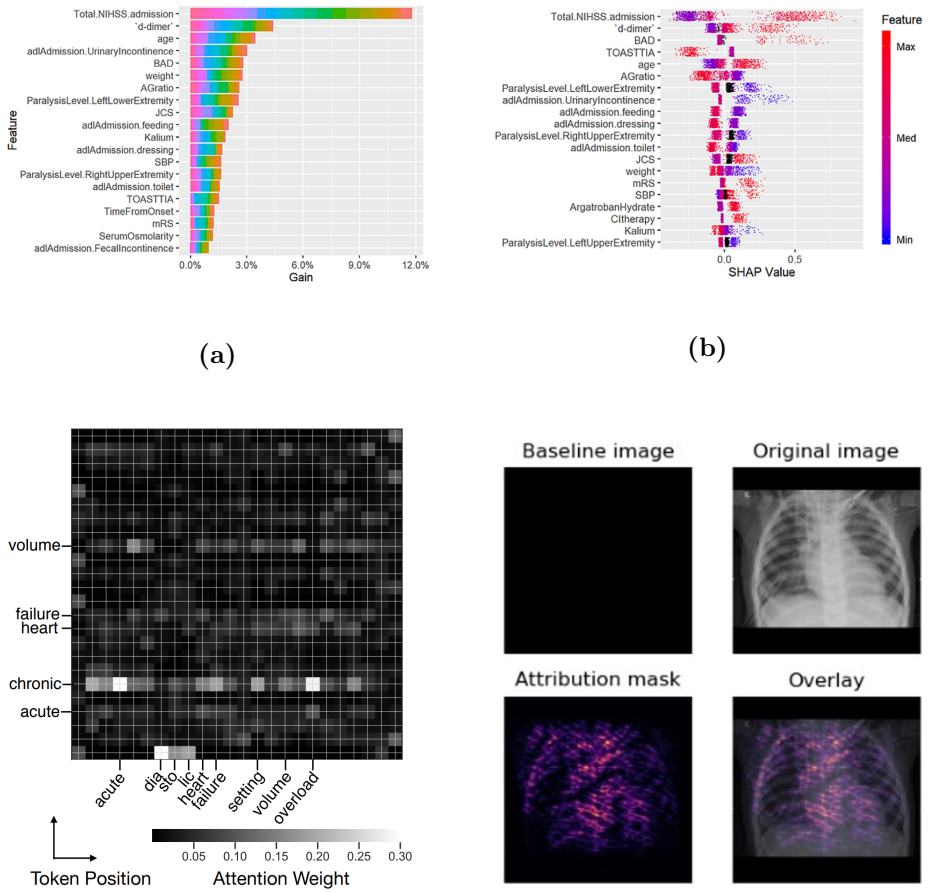

Figure 5: Methoden der AI-Erklärbarkeit: Variable Importance (a), SHAP (b), Attention Visualization (c) und Integrated Gradients (d). Reproduziert aus [27, 14, 29].

4 Was kann KI noch nicht für mich tun?

LLMs weisen eine Reihe von Einschränkungen in ihrer Anwendbarkeit auf. Im Folgenden erörtern wir diese wesentlichen Einschränkungen sowie mögliche Wege zu ihrer Überwindung.

Sie kann sich nicht immer an die Fakten halten: LLMs haben die Tendenz, "mit Sicherheit falsch" zu liegen. Dieses Phänomen, das oft als *Halluzination* bezeichnet wird, wurde in Form von Modellen beobachtet, die nicht existierende Papiere zitieren oder Details zu einer Erzählung hinzufügen, die zwar im Kontext üblich sind, aber im konkreten Fall nicht korrekt waren [21]. Halluzinationen bei der Verarbeitung klinischer Texte können dazu führen, dass ein Modell z. B. Aspekte dessen, was Patienten geäußert haben oder welche Daten zur Verfügung gestellt wurden, fabriziert, um eine Hypothese, die es zu unterstützen versucht, besser zu erfüllen. Bei der Verwendung von LLMs ist es wichtig zu bedenken, dass sie selbst keine Suchmaschinen sind und kein symbolisches Wissen enthalten [13].

Sie kann nicht unvoreingenommen sein, wenn die Daten einen Bias haben: AI-Modelle suchen im weitesten Sinne nach Mustern in den Eingabedaten, die die Ausgabe so genau wie möglich erklären. Wenn die Eingabedaten eine Verzerrung aufweisen, wird auch das Modell verzerrt [38]. Speziell bei LLMs ist ein kürzlich beobachtetes Phänomen die Modell-Schmeichelei [33]. Bei einer Patientenanamnese zum Beispiel wird das Modell seine Ausgabe verändern, wenn die Aufforderung "Ich vermute Krankheit X" enthält. In gleicher Weise kann die Antwort eines LLM auf eine medizinische Multiple-Choice-Frage leicht von richtig zu falsch (oder von falsch zu richtig) beeinflusst werden, wenn der Benutzer schreibt: "Ich bin ein [hier Fachbegriff einfügen] und die richtige Antwort ist X". Dies stellt unter anderem ein medizinisches Risiko im Falle eines Angriffs auf ein LLM dar, oder wenn die Rechtfertigung eines Modells als logische Erklärung missverstanden wird.

Sie kann das klinische Urteilsvermögen in mehrdeutigen Situationen nicht ersetzen: KI, einschließlich LLMs, ist nur begrenzt in der Lage, klinische Szenarien zu interpretieren, die von Natur aus mehrdeutig sind und ein nuanciertes Verständnis der Medizin erfordern, das über die Mustererkennung hinausgeht. Dies gilt insbesondere, wenn nicht alle Datenpunkte zur Verfügung stehen. Beispielsweise zeigt KI bei der Differenzialdiagnose oft Grenzen auf [37]. Medizinische Urteile

beruhen oft auf implizitem Wissen und Intuition, die sich durch jahrelange Erfahrung entwickelt haben, etwas, das KI nicht trivial nachbilden kann.

Sie kann kein Einfühlungsvermögen zeigen oder eine Beziehung aufbauen: KI kann zwar so programmiert werden, dass sie Muster erkennt, die auf emotionale Notlagen hindeuten, oder eine beruhigende Sprache verwendet, aber sie kann sich nicht wirklich in den Patienten einfühlen oder eine therapeutische Beziehung aufbauen, die für die Beziehung zwischen Patient und Arzt unerlässlich ist. Der KI fehlt die Fähigkeit, menschliche Emotionen auf einer persönlichen Ebene zu verstehen und nachzuempfinden, was bei sensiblen Themen wie dem Überbringen schlechter Nachrichten oder der Diskussion über Palliativpflege am Lebensende von entscheidender Bedeutung ist.

Sie lässt sich oft nicht einfach in klinische Arbeitsabläufe integrieren: Die Integration von KI in bestehende klinische Systeme und Arbeitsabläufe stellt eine weitere Herausforderung dar. Obwohl KI Daten mit beispielloser Geschwindigkeit verarbeiten und analysieren kann, ist eine sorgfältige Planung und Anpassung erforderlich, um sicherzustellen, dass diese Erkenntnisse nahtlos in die tägliche Routine von Ärzten integriert werden. Die Technologie muss mit den klinischen Protokollen übereinstimmen und den Arbeitsablauf verbessern, anstatt ihn zu stören. Dazu gehört nicht nur die technische Integration, z. B. die Kompatibilität mit klinischen Informationssystemen, sondern auch die Anpassung innerhalb des Pflege- oder Operationsteams, um sicherzustellen, dass alle Mitglieder mit diesen Instrumenten vertraut sind und sie beherrschen. Darüber hinaus wirft die Verwendung von LLMs, die nicht in das Computersystem des Gesundheitswesens integriert werden können, enorme Bedenken hinsichtlich des Datenschutzes und der Sicherheit auf. Es gibt jedoch viele Möglichkeiten, diese Einschränkung zu umgehen, denn es hat sich gezeigt, dass lokal einsetzbare Modelle mit LLMs konkurrieren können, wenn sie entsprechend trainiert werden.

Sie kann zu Dequalifizierung führen: Schließlich besteht das potenzielle Risiko, dass der Einsatz von KI zu einer Dequalifizierung (*De-skilling*) des klinischen Personals führen kann. Da KI-Systeme mehr diagnostische und prädiktive Aufgaben übernehmen, können die Fähigkeiten der Kliniker in diesen Bereichen mit der Zeit schwinden. Es ist wichtig, ein Gleichgewicht zwischen dem Einsatz von KI zur Steigerung der Effizienz und dem Erhalt wichtiger klinischer Fähigkeiten zu wahren. Es ist auch wichtig, KI als ein Werkzeug zu betrachten, das die Fachkenntnisse des Arztes verbessert und nicht ersetzt, um sicherzustellen, dass die Kliniker in den Diagnoseprozess und die Entscheidungsfindung eingebunden bleiben.

5 Zusammenfassung & Ausblick

Künstliche Intelligenz hat ein enormes Potenzial, die Thoraxchirurgie zu revolutionieren, da sie Werkzeuge bietet, die die Effizienz und Effektivität der Behandlung erheblich verbessern können. Durch die Zusammenfassung umfangreicher, komplexer Patientendaten, die Verbesserung der diagnostischen Genauigkeit, die Vorhersage von Operationsergebnissen, die Optimierung von Behandlungsplänen und die Verbesserung der postoperativen Überwachung können KI-Technologien Chirurgen in die Lage versetzen, eine bessere Versorgung bei geringerem Zeitaufwand zu gewährleisten. Allerdings ist es wichtig, die Grenzen der KI zu erkennen. Diese Technologien können ein differenziertes klinisches Urteilsvermögen, Empathie für Patienten oder die komplexen ethischen Entscheidungsprozesse in der medizinischen Praxis nicht ersetzen.

Während wir die Fortschritte der KI in der Thoraxchirurgie begrüßen, ist es wichtig, ihre Integration aus einer ausgewogenen Perspektive zu betrachten. KI sollte als leistungsstarke Ergänzung betrachtet werden, die die Fähigkeiten des Chirurgen erweitern kann, und nicht als Ersatz für das entscheidende Erfahrungswissen, das Kliniker in die Patientenversorgung einbringen. Durch den verantwortungsvollen und umsichtigen Einsatz von KI können Kliniken ihre Arbeitsabläufe und die Ergebnisse für die Patienten verbessern und gleichzeitig die Kernkompetenzen und Werte beibehalten, die ihren Beruf ausmachen.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, 2022.

[4] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[5] J. W. Ayers, N. Desai, and D. M. Smith. Regulate artificial intelligence in health care by prioritizing patient outcomes. *JAMA*, 2024.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] D. Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

[8] S. Chen, Y. Li, S. Lu, H. Van, H. J. Aerts, G. K. Savova, and D. S. Bitterman. Evaluation of chatgpt family of models for biomedical reasoning and classification. *arXiv preprint arXiv:2304.02496*, 2023.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] A. Garcia-Agundez and C. Eickhoff. When bert fails–the limits of ehr classification. *arXiv preprint arXiv:2208.10245*, 2022.

[11] K. E. Goodman, H. Y. Paul, and D. J. Morgan. Ai-generated clinical summaries require more than accuracy. *JAMA*, 2024.

[12] P. Hamet and J. Tremblay. Artificial intelligence in medicine. *Metabolism*, 69:S36–S40, 2017.

[13] W. R. Hersh. Search still matters: information retrieval in the era of generative ai. *arXiv preprint arXiv:2311.18550*, 2023.

[14] K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[15] M. M. Islam, T. Nasrin, B. A. Walther, C.-C. Wu, H.-C. Yang, and Y.-C. Li. Prediction of sepsis patients using machine learning approach: a meta-analysis. *Computer methods and programs in biomedicine*, 170:1–9, 2019.

[16] D. Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[17] N. Kroeger, D. Ley, S. Krishna, C. Agarwal, and H. Lakkaraju. Are large language models post hoc explainers? *arXiv preprint arXiv:2310.05797*, 2023.

[18] M. Krusche, J. Callhoff, J. Knitza, and N. Ruffer. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and chatgpt-4. *Rheumatology international*, pages 1–4, 2023.

[19] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[20] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[21] P. Lee, S. Bubeck, and J. Petro. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023.

[22] G. LemaÃŽtre, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18(17):1–5, 2017.

[23] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[24] B. Meskó and E. J. Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120, 2023.

[25] A. Meyer, D. Zverinski, B. Pfahringer, J. Kempfert, T. Kuehne, S. H. Sündermann, C. Stamm, T. Hofmann, V. Falk, and C. Eickhoff. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory Medicine*, 6(12):905–914, 2018.

[26] T. Nakaura, N. Yoshida, N. Kobayashi, K. Shiraishi, Y. Nagayama, H. Uetani, M. Kidoh, M. Hokamura, Y. Funama, and T. Hirai. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Japanese Journal of Radiology*, pages 1–11, 2023.

[27] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, 214:106584, 2022.

[28] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[29] J. Rabbah, M. Ridouani, and L. Hassouni. Improving pneumonia diagnosis with high-accuracy cnn-based chest x-ray image classification and integrated gradient. *Available at SSRN 4625430*.

[30] S.-E. Ryu, D.-H. Shin, and K. Chung. Prediction model of dementia risk based on xgboost using derived variable extraction and hyper parameter optimization. *IEEE Access*, 8:177708–177720, 2020.

[31] Y. Si, J. Wang, H. Xu, and K. Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, 2019.

[32] K. Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017.

[33] M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.

[34] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9):1525–1534, 2022.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] J. Vig. Bertviz: A tool for visualizing multihead self-attention in the bert model. In *ICLR workshop: Debugging machine learning models*, volume 23, 2019.

[37] Z. Yan, K. Zhang, R. Zhou, L. He, X. Li, and L. Sun. Multimodal chatgpt for medical applications: an experimental study of gpt-4v. *arXiv preprint arXiv:2310.19061*, 2023.

[38] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, 2020.

[39] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.