Tobias Schäffter, Daniel Schwabe, Stefan Haufe

Gute Daten für eine vertrauenswürdige KI in der Medizin

In:
Dössel, Olaf / Schäffter, Tobias / Rutert, Britta (Hrsg.): Künstliche Intelligenz in der Medizin.
ISBN: 978-3-949455-18-6
Berlin: Berlin-Brandenburgische Akademie der Wissenschaften, 2023
S. 49-61
(Denkanstöße aus der Akademie : eine Schriftenreihe der Berlin-Brandenburgischen Akademie der Wissenschaften ; 11)

Persistent Identifier: urn:nbn:de:kobv:b4-opus4-38069

GUTE DATEN FÜR EINE VERTRAUENSWÜRDIGE KI IN DER MEDIZIN

Tobias Schäffter, Daniel Schwabe, Stefan Haufe

Jeden Tag wird im Gesundheitssystem eine Vielzahl von Daten erhoben, die Informationen über den Gesundheitszustand von Menschen liefern, um gegebenenfalls notwendige Behandlungsschritte frühzeitig einzuleiten. Die Auswertung erfolgt derzeit durch Ärzt:innen entlang medizinischer Richtlinien. Allerdings wird das Potenzial der Daten für die Gesundheitsversorgung bisher nur im geringen Maße genutzt. Durch die schnell voranschreitende Digitalisierung stehen immer mehr Daten für solche Entscheidungen zur Verfügung. Es werden auch zunehmend Methoden der Künstlichen Intelligenz (KI) in der Medizin genutzt, um Zusammenhänge in Daten zu erfassen und Diagnosen mit zuvor gelernten Mustern zu vergleichen. KI-Methoden werden in der Forschung eingesetzt, um bisher Unbekanntes über Krankheiten zu lernen und auf dieser Basis neue Diagnose- und Behandlungsansätze zu entwickeln, die noch gezielter auf die einzelne Patientin, den einzelnen Patienten ausgerichtet sind. Beispielsweise können bevölkerungsbezogene Versorgungsdaten seltene Nebenwirkungen von Therapien aufzeigen und helfen, Patientinnen und Patienten individueller und damit besser zu behandeln. Das gilt insbesondere für komplexe Therapien, etwa bei Krebserkrankungen. KI-basierte Systeme könnten in Tausenden von Datensätzen zu vergangenen Fällen Muster erkennen, mögliche Neben- und Wechselwirkungen identifizieren und herausfinden, welche Faktoren ausschlaggebend für einen positiven Therapieansatz sein können. Neben der Verwendung von KI-Methoden in der Forschung halten diese auch zunehmend Einzug in Medizinprodukte und damit in den Arbeitsalltag behandelnder Ärztinnen und Ärzte. Die Zahl der Medizinprodukte mit KI-Anteil hat sich seit in den letzten 5 Jahren erheblich erhöht mit über 240 zugelassenen Produkten in Europa.[1] Die Mehrheit der Produkte unterstützt dabei die Arbeit von Radiologen und Kardiologen. So kann beispielsweise eine KI-unterstützte Diagnose anhand von Mustern in medizinischen Bildern oder EKG-Signalen oft schneller und präziser durchgeführt werden als durch einen einzelnen Menschen allein. Die Ärztin oder der Arzt erhalten wertvolle Hinweise, die sie mit ihrer medizinischen Erfahrung bewerten, um dann letztendlich medizinische Entscheidungen zu fällen.

1 Muehlematter UJ, Daniore P, Vokinger KN. 2021. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health. Mar; 3(3):e195e203. doi: 10.1016/S2589-7500(20)30292-2.

Im Gegensatz zu klassischen Algorithmen, bei denen der Rechenweg definiert wird, werden in der KI andere Verfahren eingesetzt, die nicht notwendigerweise zu erwartbaren Ergebnissen führen. So werden bestimmte KI-Systeme zunächst trainiert, um Zusammenhänge und Gesetzmäßigkeiten aus Daten zu lernen und diese dann auf neue Daten anzuwenden. Solche Methoden des maschinellen Lernens (ML) stellen eine Teildisziplin der künstlichen Intelligenz dar und führten in den vergangenen Jahren bereits zu sehr großen Erfolgen. Maschinelle Lernverfahren sind auch Computerprogramme (Algorithmen), deren Rechenvorschriften allerdings nicht von vornherein feststehen, sondern anhand von Trainingsdaten erlernt werden. Mit falschen oder unvollständigen Daten gespeist, können solche Verfahren der KI auch falsche oder verzerrte Ergebnisse liefern. In den meisten Fällen sind Ergebnisse des ML auch nicht nachvollziehbar. Darüber hinaus können KI-basierte Systeme problematische oder sogar diskriminierende Entscheidungen hervorrufen, etwa, wenn Trainingsdaten nur bestimmte Teile der Bevölkerung, wie z. B. junge weiße männliche Personen, repräsentieren. Es können dann unvorhersehbare und „verzerrte" Ergebnisse für andere Gruppen geliefert werden, z. B. über Ältere oder Frauen, die nicht oder nur zu geringem Teil in den Daten repräsentiert waren. Derzeit gibt es relativ wenige große Datensätze, die öffentlich zugänglich sind und bei der Entwicklung von ML-Verfahren genutzt werden können. Neben Datenschutz- und Datensicherheitsfragen zur Wahrung von Persönlichkeitsrechten spielen dabei auch wirtschaftliche Gründe eine Rolle. So werden viele medizinische Daten im Rahmen von Studien erhoben, die durch Firmen finanziert sind. Diese Daten sind dann die Geschäftsgrundlage von KI-Produkten. Daneben werden Daten, die zur Entwicklung neuer Verfahren genutzt werden, auch durch proprietäre Plattformen gesammelt. Beispielsweise hat vor Kurzem ein Team von Apple eine umfangreiche Studie zur automatischen Detektion von Herzrhythmusstörungen in einem renommierten Fachjournal veröffentlicht[2]. Dazu wurde der Herzrhythmus von mehr als 400.000 Teilnehmern über acht Monate mit einem Algorithmus der Apple Watch überwacht. Bei ca. 0.5 % der Teilnehmer konnte ein unregelmäßiger Herzschlag entdeckt werden. Diese enorme Datenmenge wurde innerhalb kurzer Zeit im Apple-Konzern nach freiwilliger Einwilligung der Nutzer gesammelt, eine Größe, wie sie bei öffentlich finanzierten Studien nur sehr selten erreicht wird. Die Daten wurden von Apple auch genutzt, um die EKG-App bei der zuständigen FDA-Behörde in den USA zuzulassen.

2 Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM, Rajmane A, Cheung L, Hung G, Lee J, Kowey P, Talati N, Nag D, Gummidipundi SE, Beatty A, Hills MT, Desai S, Granger CB, Desai M, Turakhia MP. 2019. Apple Heart Study Investigators. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. N Engl J Med. 14; 381(20): 1909–1917.

Zusätzlich wurde diese Funktion in einer klinischen Validierungsstudie an 600 Probanden getestet, wobei die Hälfte diagnostiziertes Vorhofflimmern besaß. In der Zulassung wurde klargestellt, dass das aufgenommene EKG der Apple Watch nur informieren, aber nicht diagnostizieren kann. In Europa wurde die EKG-App von Apple separat als Medizinprodukt der Risikoklasse I nach der europäischen Medizinprodukteverordnung (MDR)[3] zugelassen. Interessanterweise wurde nicht die gesamte Uhr als Medizinprodukt zertifiziert, sondern nur die Software, wobei dafür eine Selbstkonformitätsbewertung ausreichte. Das Beispiel zeigt, dass derzeit eine unabhängige Prüfung des Algorithmus und der zu Grunde liegenden Daten fehlt. Die eigentliche Zulassung erfolgte über eine klinische Studie, was im Allgemeinen recht aufwendig ist und derzeit vor allem von kleinen und mittelständischen Unternehmen als Innovationshemmnis gesehen wird. Klinische Studien bewerten im Grunde nur die Ergebnisse in einem engen Einsatzbereich der Studienkohorte und erlauben nur bedingt Aussagen über die Qualität für einen breiteren Einsatz. Es wird seit einigen Jahren an Verfahren gearbeitet, welche das Verhalten von KI-Methoden charakterisieren. Dennoch fehlen allgemein gültige Qualitätsregeln und Prüfverfahren. Um die Verwendung von KI in Medizinprodukten sicher und verlässlich zu machen und um deren Vertrauen und Akzeptanz in der Gesellschaft zu schaffen, sind neue Ansätze für eine digitale Qualitätsinfrastruktur von Nöten.

Qualität und Prüfverfahren

Die Qualität der Trainings- und Testdaten spielt eine fundamentale Rolle bei der Qualitätssicherung aller ML-Anwendungen. Wie wichtig eine gesicherte Datenlage für eine solche KI-Anwendung ist, hängt von der Kritikalität des Algorithmus ab. Der Entwurf des Artificial Intelligence Act der EU[4] etabliert hierzu vier Risikogruppen und damit einhergehende Prüfanforderungen für ML-Anwendungen. Dabei zählen Medizinprodukte und In-Vitro-Diagnostika zu den Hochrisikoanwendungen [AI Act, Punkt (30)]. Für diese gilt die Forderung nach einer „hohen Qualität der Datensätze, die in das System eingespeist werden, um Risiken und diskriminierende Ergebnisse so gering wie möglich zu halten".[5]

3 EU-Medical Device Regulation, MDR 2017/745.

4 EU AI Act: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

5 EU AI Act Richtlinien: https://ec.europa.eu/commission/presscorner/detail/de/ip_21_1682

Derzeit fehlen allerdings klar definierte Vorgehensweisen, wie solche allgemeinen Anforderungen in der Praxis umgesetzt werden können. So beklagen laut einer Umfrage der Unternehmensberatung Price Waterhouse Coopers[6] zwanzig weltweit führende Medizinproduktehersteller die „Überregulierung" als die größte Bedrohung für Innovationen. Kürzlich stellte die USA eine offizielle Anfrage bei der Welthandelsorganisation, in der die neue Medizinprodukteverordnung der EU als großes Handelshemmnis für medizinische Geräte aus den USA zum EU-Markt beschrieben wird[7], insbesondere für Geräte, die einen hohen Softwareanteil aufweisen. Es sind also klare Leitlinien und Verfahrensanweisungen für Medizinprodukte mit KI-Komponenten notwendig, um die Qualität der Daten und des KI-Verfahrens zu prüfen. Da bei der Formulierung der europäischen Medizinprodukteverordnung nur im geringen Maße an neue ML-Verfahren gedacht wurde, wird eine Anpassung erwartet. Dazu müssen Test- und Prüfverfahren für KI-Verfahren entwickelt werden, deren Nutzen wiederum in der Anwendung demonstriert werden muss, um Akzeptanz in der Gesellschaft zu schaffen.

Datenqualität

Das Grundprinzip des ML ist das Erlernen von Zusammenhängen in Daten. Aufbauend auf den gelernten inhärenten Strukturen der Daten trifft ein Algorithmus dann Vorhersagen und Entscheidungen. Die Qualität der Trainingsdaten, die zum Lernen verwendet werden, hat daher einen entscheidenden Einfluss auf die Funktionsweise und Qualität einer ML-Anwendung. Es gibt derzeit eine intensive Diskussion darüber, dass der Einsatz von ML zum Teil zu falschen oder diskriminierenden Entscheidungen führen kann. Dies liegt weniger am Algorithmus als an fehlerhaften oder unvollständigen Trainingsdaten, die die Qualität beeinflussen. Aus diesem Grunde ist eine Prüfung und Absicherung der Datenqualität unerlässlich. Beispielsweise sieht die zuständige Behörde in den USA die Wahl der Trainings- und Testdaten als Schlüsselkomponente für den erfolgreichen Einsatz von ML-Techniken. Daten sollten daher unter Verwendung von Qualitätssicherungs- und -Managementsystemen erhoben werden.[8] Dies beinhaltet u. a. Protokolle zur Datenerhebung, Bestimmung eines Referenzstandards sowie die Auditierung von

6 Price Waterhouse Coopers; 20th CEO Survey: Healthcare industry key findings 2017.

7 World Trade Organization (G/TBT/W/679); Statement by the USA to committee on technical barriers to trade; Juli 2019 link: G/TBT/W/679 24 July 2019 (19-4907) Page

8 FDA "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning [AI/ML] Based Software as a Medical Device [SaMD]"

Test- und Trainingsdaten innerhalb der Hersteller-Organisation. Allerdings ist derzeit keine unabhängige Prüfung notwendig. Auch in Europa wird im Rahmen des „Artificial Intelligence Act“ [9] eine Ausgestaltung von Verfahren zur Datenqualität für KI-Verfahren in der Medizin gefordert. Dabei stehen die Qualitätsmerkmale wie Richtigkeit, Präzision, Vergleichbarkeit, Vollständigkeit und Repräsentativität von Daten im Vordergrund.

Damit eine ML-Anwendung die „richtigen“ Entscheidungen treffen kann, müssen die Trainingsdaten eine hohe Genauigkeit aufweisen. Die Genauigkeit ist eine zentrale Fragestellung der Metrologie, der „Wissenschaft des Messens“, wobei hier zwischen Richtigkeit und Präzision unterschieden wird. Die Richtigkeit (oder Fehler, „Bias“) beschreibt die Nähe (oder Ferne) eines Datenwertes zu einem „wahren“ Wert (meist einem Referenzwert). Dagegen beschreibt die Präzision die Abweichung der Daten untereinander, d. h. wieweit die Datenwerte um einen Mittelwert streuen. Systematische Abweichungen können entstehen, wenn beispielsweise unterschiedliche Messsysteme verwendet werden, deren Messungen im Mittel streuen und eine unterschiedliche Nähe zum wahren Wert haben. Beide Qualitätsmerkmale haben eine große Bedeutung für die Entwicklung von ML-Verfahren. Insbesondere führt ein großer systematischer Fehler („Bias“) zu einem verzerrten Lernergebnis, so dass in der Anwendung des Erlernten, neue Werte konsistent falsch eingeordnet werden. Üblicherweise soll aber eine Abweichung von einem Referenzwert zur Diagnose einer krankhaften Veränderung genutzt werden. Daher ist es wichtig zwischen der systematischen Abweichung aufgrund des Messsystems und der Abweichung vom „Normwert“ aufgrund einer Krankheit zu unterscheiden. Das gilt auch für die Präzision; auch hier muss zwischen einer Streuung der Messungen und der biologischen Varianz unterschieden werden. Der Präzision kommt auch eine wichtige Rolle zu, wenn es darum geht, zuverlässige Entscheidungsschwellwerte festzulegen. Oft werden Daten aus unterschiedlichen Quellen und unter Verwendung unterschiedlicher Messverfahren zusammengeführt. Dabei kommt es zu einer Mischung der verschiedenen Präzisions- und Genauigkeitswerte, welche selbst ein Muster bilden können. Falls diese Herkunft nicht berücksichtigt wird, kann dieses „Muster“ ungewollt erlernt werden und es kann die eigentlichen Merkmale, jene die mit der eigentlichen Krankheit zu tun haben, überdecken. Dies ist vor allem der Fall, wenn sogenanntes „confounding“ vorliegt, z. B. wenn sich die relative Häufigkeit von Diagnosen zwischen Datenquellen unterscheidet. Auf Testdaten angewandt, würde das trainierte ML-Verfahren dann, anstatt die gewünschte Klassifikation der Krankheit auf Basis klinisch-

9 https://artificialintelligenceact.eu, 2021.

relevanter Dateneigenschaften durchzuführen, vorrangig die geschätzte Herkunft (d. h. das Messverfahren) verwenden. Neben der Genauigkeit der Daten aufgrund der ursprünglichen Messung, ist die „ungenaue" Zuordnung der Daten (sog. „Label") durch Experten eine weitere Fehlerquelle beim ML.

Die Vergleichbarkeit von Daten beruht auf der Verwendung eines einheitlichen Referenzsystems. Die beste Möglichkeit, die Vergleichbarkeit von Daten – unabhängig davon, wann und wo sie ermittelt wurden – zu gewährleisten, ist die metrologische Rückführung auf ein gemeinsames Referenzsystem. Mit dem Internationalen Einheitensystem (SI) gibt es ein solches System. Es bildet seit der Meterkonvention von 1875 auch die weltweite Grundlage für Handel in über neunzig Staaten. Das SI-System ist fester Bestandteil in vielen Bereichen des täglichen Lebens. Allerdings ist das SI-System im Gesundheitswesen noch nicht vollständig verbreitet. Es sind große EU-weite Initiativen vonnöten, um die Vergleichbarkeit von medizinischen Messverfahren durch metrologische Rückführung weiter zu verbessern und so auch die Qualität von multizentrischen Studiendaten zu erhöhen.

Repräsentativität und Vollständigkeit sind weitere statistische Qualitätsmerkmale für Daten mit hoher Bedeutung für das ML. Dies umfasst zum Beispiel das Verhindern von Diskriminierung jeglicher Form durch eine statistische Unterrepräsentation bestimmter Datengruppen innerhalb der Trainingsdaten. Damit wird gewährleistet, dass ein ML-Algorithmus lernen kann, richtige Entscheidungen und Vorhersagen für alle vorgesehenen Anwendungsbereiche zu treffen. Neben der Qualität der Trainingsdaten spielt die Qualität der Testdaten ebenfalls eine wesentliche Rolle. Diese werden verwendet, um neue ML-Verfahren sowohl zu testen als auch zu verbessern. Dabei gelten die gleichen Qualitätskriterien wie für Trainingsdaten, allerdings mit höherer Bedeutung einzelner Kriterien. Beispielsweise ist die Repräsentativität der Testdaten entscheidend dafür, die Eignung einer ML-Anwendung für unterschiedliche Szenarien gewährleisten zu können. Wesentlich ist hierbei auch, die Trainingsdaten streng von den Test- und Validierungsdaten zu trennen.

Qualität des ML-Verfahrens

ML-Verfahren werden anhand von Daten trainiert, wobei die Entscheidungsregeln vom System selbst erlernt werden, ohne dass diese Zusammenhänge direkt sichtbar sind. Daher wird häufig von einer „Black Box" gesprochen. Oft wird der Wunsch nach Veröffentlichung und Transparenz der Verfahren geäußert. Dies

greift allerdings recht kurz, da selbst bei der Veröffentlichung aller Werte (sog. Gewichte) eines neuronalen Netzwerkes das Verhalten zwar reproduziert werden kann, dieses oft aber selbst vom Entwickler nicht vollständig verstanden wird. Insgesamt ist der Einfluss der Trainingsdaten auf das Verhalten im maschinellen Lernen so stark, dass eine komplett unabhängige Prüfung der beiden Einzelaspekte nicht sinnvoll und möglich ist. Derzeit werden Kriterien und Metriken für die Beurteilung der Qualität von ML-Verfahren zur Validierung entwickelt. Gerade da medizinische Entscheidungen kritisch sein können, sollte es klar definierte Kriterien geben. Dabei stehen die Qualitätsmerkmale wie Leistungsfähigkeit, kalibrierte Unsicherheitsquantifizierung, Erklärbarkeit, Generalisierbarkeit und Robustheit im Vordergrund.

Die Vorhersagegüte (Performance) ist eines der wichtigsten Kriterien zur Beurteilung von ML-Verfahren. Je nachdem, ob die Zielgrößen kontinuierlich oder kategorial sind, kommen neben klassischen metrologischen Fehlermaßen auch Maße aus der Signaltheorie wie Sensitivität und Spezifizität zum Einsatz. Zur Bestimmung der Vorhersagegüte ist es allerdings notwendig, die „richtigen" Werte (Referenzwerte) zu kennen. Daneben ist aber auch die Unsicherheit von Vorhersagen von ML-Modellen gerade in der Medizin von hoher Bedeutung. Hier geht es darum, dem/der klinischen Entscheidungsträger:in nicht nur einen einzelnen Wert an die Hand zu geben, sondern auch eine Einschätzung, wie sicher diese Vorhersage ist. Wenn die Unsicherheit zu hoch ist, kann ein Arzt oder eine Ärztin dies im Entscheidungsprozess berücksichtigen. Dabei werden zwei Haupttypen der Unsicherheit unterschieden. Die epistemische Unsicherheit beschreibt, was das Modell nicht wissen kann, weil die Trainingsdaten nicht angemessen, unvollständig oder in ihrer Anzahl nicht ausreichend sind oder weil die Komplexität des Modells zur Modellierung der Daten nicht ausreicht. Demgegenüber bezieht sich die aleatorische Unsicherheit auf die inhärente Zufälligkeit der Daten, wobei der Begriff *Aleator* im Lateinischen jemanden beschreibt, der würfelt. Bei genügend Trainingsdaten nimmt die epistemische Unsicherheit ab, während die aleatorische Unsicherheit nicht verringert werden kann, selbst wenn mehr Daten bereitgestellt werden. Insgesamt gibt es für die Bestimmung der Unsicherheit von Entscheidungen verschiedene Ansätze. Klassische Bottom-up-Ansätze der Metrologie werden im GUM-Framework[10] beschrieben. Die grundlegende Idee ist hierbei, dass bekannte Verteilungen und Unsicherheiten der Eingangsgrößen benutzt werden können, um entsprechende Verteilungen und Unsicherheiten der Ausgangsgrößen

10 BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML. 2008. Guide to the Expression of Uncertainty in Measurement. JCGM 100:2008, GUM 1995 with minor corrections.

einer Funktion zu schätzen. Dieser Ansatz setzt aber voraus, dass die angewandte Funktion statisch ist und nicht von den Daten selbst abhängt. Dies ist im ML jedoch nicht immer gegeben. Stattdessen kann man das Schätzen eines ML-Modells als inverses Problem auffassen, bei der die Verteilung der Modellparameter aus den beobachteten Daten geschätzt wird. Aus der Verteilung der Parameter kann dann wiederum die Verteilung der Vorhersagen abgeleitet werden. Der Nachteil dieser Modellierung ist jedoch eine hohe Subjektivität, da es notwendig ist, Vorwissen über die Verteilungen der Modellparameter zu spezifizieren. Dieser Nachteil kann bei guter Datenlage jedoch zum Teil durch sogenannte empirische Bayes-Verfahren ausgeglichen werden. Ein weiteres interessantes Paradigma zur Unsicherheitsschätzung ist die Top-down-Modellierung. Hierbei wird beispielsweise die Verteilungsfunktion möglicher Ergebnisse aus den Daten gelernt. Ein Beispiel sind „Monte Carlo Dropout"-Ansätze[11] für Deep Learning, welche die Unsicherheit dadurch bestimmen, dass beim Trainieren einzelne „Neuronen" mit einer gewissen Wahrscheinlichkeit deaktiviert (sog. „Dropout") werden, wodurch unterschiedliche Vorhersagen erzielt werden. Nach verschiedenen Durchläufen und unterschiedlichen Deaktivierungen kann dann eine Verteilung der Ergebnisse und daraus eine Unsicherheit bestimmt werden. Ein weiteres Beispiel sind Methoden, die Unsicherheitsparameter (z. B. Standardabweichung oder Perzentile) direkt schätzen, wie sogenannte „deep ensembles".[12] Dafür werden dem neuronalen Netz weitere Neuronen in der Ausgabeschicht hinzugefügt. Diese Modellierung ist komplett datengetrieben. Dabei werden die gelernten Unsicherheitsintervalle durch geeignete Verlustfunktionen (sogenannte „proper scoring functions") kalibriert, sodass die angegebene Überdeckung zumindest bei der Trainingsstichprobe auch tatsächlich zutrifft. Eine weitere Frage ist, wie Unsicherheitswerte von unterschiedlichen neuronalen Netzen verglichen werden können.

Ziel der „Erklärbarkeit" ist es, die vom ML-Verfahren erlernten Zusammenhänge in den Daten zu kennzeichnen. In den vergangenen Jahren wurden Methoden entwickelt, die den menschlichen Nutzerinnen und Nutzern aufzeigen sollen, wie der ML-Algorithmus zu seiner Entscheidung gekommen ist. Dieser Aspekt ist von zentraler Bedeutung in ML-Anwendungen zur Unterstützung von Medizinern, welche maschinelle Ergebnisse verstehen wollen, um so Vertrauen für eine Entschei-

11 Gal Y, Ghahramani Z. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning; Proceedings of the 33rd International Conference on International Conference on Machine Learning. 48: 1050–1059.

12 Lakshminarayanan B., Pritzel A., & Blundell C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30.

dung auf dieser Grundlage zu gewinnen. Das Forschungsgebiet der „erklärbaren KI" (engl. Explainable Artificial Intelligence – XAI) wurde mit großen Forschungsprogrammen in den letzten Jahren sehr stark entwickelt (beispielweise wurden in den USA 70 Millionen Dollar für das „Explainable-AI"-Forschungsprogramm[13] zur Verfügung gestellt). Oft zielt „erklärbare KI" darauf ab, Charakteristika in Daten zu identifizieren, welche einen besonders starken Einfluss auf die getroffene Entscheidung des Verfahrens hatten[14]. Diese Datenbereiche können dann z. B. farbig markiert werden, um die Information für den Benutzer und die Benutzerin sichtbar darzustellen. Eine wichtige Technik ist das „Layer-wise Relevance Propagation" LRP-Verfahren[15], das schrittweise die Entscheidungswerte bestimmt. Allerdings wird bezweifelt, ob der „Einfluss" einer Variablen allein ausreicht, um belastbare Aussagen über das Wirkungsprinzip eines trainierten ML-Modells generell oder auf einem konkreten Datenpunkt zu verstehen. So können sogenannte „Suppressorvariablen" mit Störsignalen beispielsweise einen starken Einfluss auf die Vorhersage haben, ohne selbst irgendeine statistische Abhängigkeit zum Vorhersageziel aufzuweisen[16]. Insgesamt ist festzustellen, dass es im Feld der „erklärbaren" KI derzeit keine ausreichend mathematisch fundierten Definitionen von Korrektheit gibt, mit Hilfe derer die „Erklärgüte" von XAI-Methoden objektiv bewertet werden könnte.

Die Generalisierbarkeit beschreibt die Eigenschaft eines ML-Verfahrens, für möglichst viele verschiedene Eingabedaten und Anwendungsszenarien gültige Ausgaben zu liefern. Dieses Ziel muss während des Trainierens eines Modells berücksichtigt werden, indem eine Überanpassung an die Trainingsdaten verhindert wird. Danach sollten die trainierten Modelle an verschiedenen Daten außerhalb des Trainingsdatensatzes getestet werden, um ihre Generalisierbarkeit bewerten zu können. Beispielsweise kann durch eine gezielte Wahl von Testdaten, die sich in einigen ihrer Eigenschaften von den Trainingsdaten unterscheiden (sog. „out-of-distribution data"), die Generalisierbarkeit einer ML-Anwendung untersucht werden. Während Generalisierbarkeit die Fähigkeit von Modellen ist, Datenpunkte außerhalb der Trainingsdaten vorherzusagen, bezieht sich die Robust-

13 Voosen P. 2017. The AI detectives. Science. 357(6346): 22–27.

14 Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. 2019. Unmasking clever hans predictors and assessing what machines really learn. Nature communications, 10(1): 1–8.

15 Montavon G, Samek W, Müller KR. 2017. Methods for Interpreting and Understanding Deep Neural Networks Digital Signal Processing, 73: 115.

16 Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J D, Blankertz B & Bießmann, F. 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage, 87: 96–110.

heit auf die Stabilität eines ML-Verfahrens gegenüber äußeren und feindlichen Einflüssen. Das Kriterium „Robustheit" wird meist in der Softwaresicherheit verwendet und beschreibt die Eigenschaft eines ML-Systems, gegenüber Angriffen durch falsche Daten zu bestehen. Schon im Jahr 2014 haben Forscher von Google die Anfälligkeit bestimmter neuronaler Netze nachgewiesen und konnten Daten so manipulieren, dass ML-Verfahren falsche Entscheidungen fällten. Solche feindlichen („adverserial") Daten werden im Normalfall vom menschlichen Betrachter nicht erkannt, da sie oft nur auf minimalen Änderungen der Originaldaten basieren. Sie führen aber zu falschen ML-Entscheidungen, da sie zu Angriffszwecken konstruiert wurden. Die entwickelten Verteidigungsmethoden können verwendet werden, um ML-Verfahren, die über ihre Eingabemodalitäten Risiken zu sicherheitsrelevanten Manipulationen des Systemverhaltens bieten, abzusichern und mit geeigneten Verfahren die Robustheit des Systems zu messen. Eine Möglichkeit die Robustheit zu erhöhen, ist die Verwendung von Filtern, welche manipulative Veränderungen reduzieren. Es können auch informationstheoretische Methoden verwendet werden, um zu prüfen, ob Daten weitere potenziell feindliche Informationen enthalten.[17]

Vergleichstests

Neben den genannten Kriterien wird noch an weiteren Eigenschaften gearbeitet, um die Qualität von ML-Verfahren möglichst umfassend bewerten zu können. Solche Kriterien bilden die Grundlage für Vergleiche („Benchmarktests") und erlauben die Definition von Referenzverfahren mit definierter Qualität, an denen sich neue Entwicklungen messen können. Allerdings gibt es aufgrund der Vielfalt von Anwendungen keine allgemein gültige Qualitätseinschätzung, sondern eher Richtlinien. Da die Qualität der ML-Verfahren maßgeblich von der Datenqualität abhängt, ist ein fairer Vergleich nur unter Verwendung gleicher Validierungsdaten möglich. Zur Beurteilung maschineller Lernverfahren werden regelmäßig Wettbewerbe (sog. „Challenges") ausgetragen, in denen die Leistungsfähigkeit der ML-Modelle anhand vorgeschriebener Bewertungskriterien mit standardisierten Datensätzen verglichen werden. In diesem Zusammenhang hat die Definition von Referenzdaten einen hohen Stellenwert. Auch wenn es nicht möglich ist, Referenzdaten für alle Anwendungsfälle zur Verfügung zu stellen, können diese für bestimmte Klassen wie Einzelwertmessungen (z. B. Temperatur, Blutdruck, Sauer-

17 Martin J, Elster C. 2020. Inspecting adversarial examples using the fisher information. Neurocomputing, 382. 80–86.

stoffsättigung), Zeitreihen (Elektrokardiogramm, Pulswellen) oder medizinische Bilder für verschiedene Fragestellungen definiert werden. Leider ist das Angebot an offenen Referenzdatensätze gering bzw. die Größe der Datensätze so klein, dass ein guter Vergleich nicht immer möglich ist. Es gibt aber große internationale Initiativen mit dem Ziel, dies zu verbessern. Beispielsweise wurde kürzlich ein offener Datensatz mit über 21.000 Elektrokardiogramm (EKG)-Messungen von mehr als 18.000 Patienten und Gesunden veröffentlicht.[18] Jede klinische EKG-Messung wurde von 10 Elektroden erfasst und danach von zwei Kardiologen diagnostiziert und entlang 71 standardisierter Klassen eingeordnet. Der Datensatz weist eine gute Verteilung sowohl zu verschiedenen Krankheitsklassen als auch zu Gesunden auf. Darüber hinaus sind auch verschiedene demografische Merkmale (z. B. Alter und Geschlecht) repräsentiert. Neben der diagnostischen Qualität gibt es auch Hinweise zur unterschiedlichen EKG-Signalqualität. Schwerpunkt des Datensatzes ist die Definition von Unterdatensätzen für Training, Test und Validierung mit möglichst einheitlicher Repräsentativität. Dabei wurde insbesondere auf eine hohe Label-Qualität bei den Test- und Validierungsdaten gelegt. Der Datensatz wurde kürzlich für eine Vergleichsstudie verwendet[19], um bestehende ML-Verfahren entlang definierter Kriterien (Leistungsfähigkeit, Unsicherheit, Erklärbarkeit) zu untersuchen und so einen strukturierten Ansatz für zukünftige Vergleiche und eine Einordnung neuer Verfahren zu etablieren. Eine potenzielle Fehlerquelle von Vergleichstests anhand solcher klinischer Datensätze ist, dass Fehler bei der Beurteilung durch Mediziner nicht ausgeschlossen werden können und nur durch erheblichen Aufwand, d. h. Beurteilung der Daten durch möglichst viele Experten, minimiert werden können. Ein alternativer Ansatz ist die Verwendung von synthetischen Daten. Diese simulierten Messdaten werden beispielsweise durch biophysikalische Modelle anhand wohl definierter Parameterwerte erzeugt und erlauben daher eine bessere Zuordnung und exakte Berechnung von Fehlern. Im Rahmen des EU-Projektes Medalcare[20] wurden dazu EKG-Daten einer virtuellen Population erstellt.[21] Solche simulierten Daten erlauben es, sowohl die Präzision als auch die Genauigkeit zu ändern, so dass der Einfluss der Datenqualität auf ein ML-Verfahren untersucht werden kann und auf diese Weise Vorgaben

18 Wagner P, Strodthoff N, Bousseljot R.-D., Kreiseler D, Lunze F, Samek W, Schaeffter T. 2020. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 7: 154.

19 Strodthoff N, Wagner P, Schaeffter T, Samek W. 2020. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL, IEEE J Biomed Health Inform.

20 EU-EMPIR Projekt „Medalcare". 2019. https://www.ptb.de/empir2019/medalcare/home/

21 Nagel C, Schuler S, Dössel O, Loewe A. 2021. A bi-atrial statistical shape model for large-scale in silico studies of human atria: Model development and application to ECG simulations. Med Image Anal.74:102210. doi: 10.1016/j.media.2021.102210.

für ein Mindestmaß an Messdatenqualität festgelegt werden können. Besonders interessant ist es, bewusst bestimmte Daten im Training wegzulassen, zu verändern bzw. eine falsche Zuordnung (Label) der Daten durchzuführen, um so die Generalisierbarkeit und Robustheit eines Verfahrens zu untersuchen. Neben dem genannten EKG-Beispiel gibt es viele weitere Initiativen, insbesondere in der biomedizinischen Bildgebung.

Zusammenfassung

Verfahren der künstlichen Intelligenz beruhen auf Daten und Algorithmen. Für beide sind neue Ansätze in der Qualitätssicherung notwendig, um die Zertifizierung und den Einsatz von KI-Verfahren zu beschleunigen und gleichzeitig ein Vertrauen in der Gesellschaft zu fördern. Da die Qualität der Verfahren wesentlich von der Datenqualität abhängt, braucht es Richtlinien für eine objektive Bewertung von Datensätzen. Dies gilt sowohl für öffentliche als auch für private Daten, bzw. Daten, die aus Geschäftsmodellgründen in Unternehmen verbleiben müssen. Richtlinien zur Datenqualität können auch für eine Kuratierung von klinischen Daten zu Referenzdatensätzen genutzt werden, welche sowohl in der Forschung als auch in der Produktentwicklung eingesetzt werden sollten. Insgesamt brauchen Medizinprodukte mit ML in Zukunft eine zusätzliche ML-bezogene Konformitätsbewertung. Hierzu sollten sowohl Trainingsdaten als auch Testdaten von unabhängigen Stellen bewertet werden. Derzeit werden solche regulatorischen Ansätze von Herstellern oft als Hürde empfunden, mittelfristig können diese aber zu einem Wettbewerbsvorteil insbesondere gegenüber den USA und China führen. Eine Bewertung der Qualität der verwendeten Daten und der entwickelten Verfahren sollte auch Auswirkungen auf die weiterhin notwendigen klinischen Studien haben. So könnten klinische Studien kleiner ausfallen, wenn eine hohe Qualität der Trainingsdaten nachgewiesen wurde, was letztendlich zu einer schnelleren Markteinführung führen kann. Daneben schafft ein Qualitätssiegel auch Vertrauen bei Kundinnen und Kunden und führt somit zu einem potenziellen Wettbewerbsvorteil gegenüber Produkten und Dienstleistungen anderer Anbieter ohne Qualitätssiegel. Das Label „Made in Germany" bzw. „Made in Europe" würde als hoher Qualitätsstandard so auf das digitale Zeitalter übertragen werden, um eine Vertrauensbildung beim Verwender und eine Erhöhung des Wettbewerbsvorteils von Herstellern zu fördern. Es ist ein wirtschafts- und gesellschaftspolitischer Rahmen notwendig, der durch folgende Aktivitäten unterstützt werden sollte:

- Entwicklung von europäischen Zulassungsvorschriften für eine zuverlässige und vertrauenswürdige KI in der Medizin.

- Entwicklung europäischer Standards und Normen für Daten- und KI-Qualität.

- Schaffung sog. Reallabore, um in Ermangelung verbindlicher regulatorischer Vorschriften bereits gemeinsam Richtlinien zwischen Herstellern, Behörden, Benannten Stellen, medizinischen Experten und Patientinnen und Patienten zu erarbeiten. Dies beinhaltet die Definition von notwendigen Qualitätskriterien sowohl für Datensätze für ML-Verfahren als auch für ML-Verfahren.

- Interdisziplinäre Zusammenarbeit von Experten an medizinischen Zentren, um hochqualitative Datensätze zu generieren.

- Bereitstellung von Referenzdaten und Vergleichstests für Forschung und Entwicklung.

- Vertrauenswürdige Prüfung von nicht-öffentlich zugänglichen Daten und ML-Verfahren durch unabhängige Stellen und Vergabe von Qualitätssiegeln.

- Regeln für notwendige klinische Studien zum Nachweis der Wirksamkeit eines ML-Systems.

- Erhöhung der gesellschaftlichen Akzeptanz gegenüber vertrauenswürdigen KI-Verfahren durch qualitätsgesicherte KI.

- Motivation zur Datenspende, um große Datensätze für die KI-Forschung zum Nutzen aller zur Verfügung stellen zu können.