MedWiss.Online

Unfaire Algorithmen benachteiligen Minderheiten

Künstliche Intelligenz hat in der Medizin bereits Einzug gehalten: Smarte Algorithmen unterstützen Ärztinnen und Ärzte bei der Diagnose. Auch in den Neurowissenschaften kann selbstlernende Software allein anhand von Gehirnscans Rückschlüsse auf persönliche Eigenschaften zulassen – auf mentale Fähigkeiten, Verhaltensweisen oder Charaktermerkmale. Doch dass diese Algorithmen eine systematische Verzerrung, einen sogenannten Bias, hinsichtlich ethnischer Minderheiten aufweisen können, hat ein internationales Forschungsteam kürzlich im Fachmagazin Science Advances statistisch bewiesen.

KI-Modelle könnten in den Neurowissenschaften dabei helfen, psychische Erkrankungen frühzeitig zu diagnostizieren. „Darin werden im Moment große Hoffnungen gesetzt“, erklärt die Erstautorin der Studie, Dr. Jingwei Li vom Jülicher Institut für Neurowissenschaften und Medizin (INM-7). „Noch sind die Algorithmen nicht reif für die Praxis. Aber große Firmen wie Google und Facebook investieren in die Forschung mit Gesundheitsdaten für die Präzisionsmedizin. Unser Anliegen war es, jetzt schon auf die Probleme hinzuweisen, die mit diesen Anwendungen verbunden sind.“

Eines der Probleme: Offenbar sind die Vorhersagemodelle anfällig für statistische Verzerrungen. Solch einen Bias konnte das Team unter Jülicher Leitung nachweisen: Ein Algorithmus, der aus Gehirnscan-Aufnahmen auf individuelle persönliche Eigenschaften schloss, lag bei Afroamerikanern häufiger daneben als bei weißen US-Bürgern.

„Wir haben öffentlich zugängliche Bilddaten von Gehirnscans ausgewertet“, sagt Prof. Dr. Sarah Genon, die am INM-7 die Arbeitsgruppe Kognitive Neuroinformatik leitet und an der Publikation federführend beteiligt war. Die Daten stammen von zwei wissenschaftlichen Großprojekten, die in den USA erhoben wurden. Mit Hilfe der funktionellen Magnetresonanztomographie wurden dafür in den Gehirnen von Probanden Konnektivitätsmuster aufgezeichnet. Zu jeder Aufnahme existieren in den Datenbanken auch weitere Angaben: zum Alter, Geschlecht und zu Daten aus psychologischen Tests der Versuchspersonen.

„Einen Großteil der Daten haben wir genutzt, um unser Modell zu trainieren“, erklärt Prof. Dr. Sarah Genon. „Mit Methoden des maschinellen Lernens lassen sich darin subtile Muster aufspüren, die beide Informationen miteinander verknüpfen – die Konnektivitätsmuster des Gehirns mit den individuellen psychischen Merkmalen der Probanden. Auf diese Weise können wir dann die Eigenschaften von Menschen vorhersagen, von denen wir nur die Gehirnscans kennen.“

An den übrigen, nicht zum Training verwendeten Daten konnte das Team anschließend überprüfen, wie treffsicher die Prognosen des Vorhersagemodells wirklich waren. Dabei zeigte sich, dass die Präzision für weiße Amerikaner höher war als für Afro-Amerikaner. Wobei es egal war, ob sich aus den Gehirnbildern grundlegende Dimensionen der Persönlichkeit herauslesen ließen, wie etwa die Offenheit für Erfahrungen, oder geistige Eigenschaften, zum Beispiel die kognitive Flexibilität. In den meisten Fällen waren die Vorhersagen für Afroamerikaner mit einem größeren Fehler behaftet.

„Wir sprechen hier von einem ‚unfairen‘ Modell. Da es für eine bestimmte Minderheit weniger verlässliche Ergebnisse liefert, könnten sich für die Angehörigen dieser Gruppe Nachteile ergeben, etwa bei einer psychotherapeutischen Behandlung“, sagt Dr. Jingwei Li. Und Prof. Dr. Sarah Genon

ergänzt: „Die Ursache hierfür liegt zum Teil in den Daten an sich, die üblicherweise für solche Analysen genutzt werden und mit denen auch wir unser Modell gefüttert haben. Denn die Bilder der Gehirnscans in den großen Datenbanken stammen überwiegend von der Mehrheit der weißen US-Bevölkerung. Daten von Afroamerikanern sind unterrepräsentiert."

Doch alleine an der Zusammensetzung der Trainingsdaten kann es nicht liegen, dass das Modell weniger akkurate Prognosen für Afroamerikaner liefert. Denn im nächsten Schritt nutzte das Team ausschließlich Trainingsdaten, die an afroamerikanischen Probanden erhoben wurden. Damit sollte sich das KI-Modell besser an diese Gruppe anpassen und die relevanten Merkmale effizienter aus den Datenbergen extrahieren. Tatsächlich verbesserte sich die Vorhersagegenauigkeit des Algorithmus. Allerdings nicht so stark wie erwartet, sagt Dr. Jingwei Li: „Obwohl der Vorhersage-Bias für Afroamerikaner abgenommen hat, war er immer noch höher als bei weißen Amerikanern."

Der Grund dafür, vermutet Prof. Dr. Sarah Genon, könnte in der Aufbereitung der Daten liegen: „Die Daten aus der Bildgebung werden von uns nach einem gewissen Schema vorbehandelt, um sie miteinander vergleichbar zu machen. Das ist ein generell notwendiger Schritt, der immer bei diesen Analysen durchgeführt werden muss. Dieses Schema und alle anderen Vorbereitungsschritte wurden basierend auf einer Bevölkerung entwickelt, die in der Mehrheit aus Menschen mit europäischer Abstammung besteht. Auch das dürfte zu dem Bias beitragen, den wir in unserer Veröffentlichung beschreiben."

Um auf Basis des maschinellen Lernens faire Modelle zu entwickeln, sei darüber hinaus eine größere Diversität der Trainingsdaten als bisher erforderlich, folgert Dr. Jingwei Li. Minderheiten müssten darin stärker vertreten sein, führt Prof. Dr. Sarah Genon an und führt weiter aus: „Die englischen Begriff ‚Race' beziehungsweise ‚Ethnicity', die wir in unserer Veröffentlichung benutzen, sind äußerst vielschichtig. Es kommt auch darauf an, welcher Kultur und welcher Gemeinschaft sich die untersuchten Menschen zugehörig fühlen. Historische, gesellschaftliche und erzieherische Faktoren spielen dabei eine wichtige Rolle."

Gerechte Algorithmen, die Minderheiten nicht benachteiligen, sind also prinzipiell möglich. Doch bevor solche Vorhersagemodelle in der Praxis Verwendung finden, sollte die Grundlage für ihre Trainingsdaten noch weiter verbessert werden.

Originalpublikation: Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity, Li et al., Sci. Adv. 8, eabj1812 (2022), DOI: 10.1126/sciadv.abj1812