Datenschutz-Risiken in Medizin-KI entdeckt

KI-Modelle, etwa für die Erkennung von Krebs, werden mit Gesundheitsdaten von Patientinnen und Patienten trainiert. Bereits die Information, dass persönliche Daten in ein Modell eingeflossen sind, kann negative Auswirkungen für die Betroffenen haben, wenn sie in die falschen Hände gerät. Im Fachmagazin "Nature" hat ein Forschungsteam jetzt gezeigt, dass den Modellen diese sensiblen Informationen mit der passenden Methode deutlich effektiver entlockt werden können als bislang gedacht.

Die Forschenden der Technischen Universität München (TUM), des Imperial College London und des Hasso-Plattner-Instituts (HPI) konnten zeigen, dass bisherige Berechnungen zur Sicherheit von KI-Modellen irreführend sind. Angriffe, mit denen sich herausfinden lässt, ob Daten einer Person in ein Modell eingeflossen sind, werden in der Fachsprache als Membership Inference Attacks (MIAs) bezeichnet. Bislang galten gängige Medizin-KI-Modelle als weitgehend sicher vor MIAs.

"Leider haben die entsprechenden Tests immer nur ein durchschnittliches Risiko für alle Patienten ermittelt. Wir haben uns erstmals das Risiko für individuelle Patienten angeschaut - für diese zeigt sich ein ganz anderes Bild", sagt TUM-Wissenschaftler Moritz Knolle, Erstautor der Studie. Während die Angriffe für einen großen Teil der Datensätze erfolglos waren, konnten manche Patientinnen und Patienten mit nahezu hundertprozentiger Wahrscheinlichkeit den Modellen zugeordnet werden. "Das ist kein verschmerzbares Risiko. Gesundheitsdaten sind hochsensibel", sagt Daniel Rückert, Professor für KI und Informatik in der Medizin an der TUM und gemeinsam mit Prof. Georg Kaissis (HPI) Letztautor der Studie.

Unterschiedliche Datentypen berücksichtigt

Die Forschenden attackierten Modelle, die auf sieben etablierten Medizindatensätzen basierten. Grundlage für die Modelle waren jeweils unterschiedliche Datentypen wie Bildgebungsdaten, Kardiogramme oder elektronische Patientenakten. "Ein Angreifer braucht drei Dinge, um eine MIA auszuführen", erläutert Georg Kaissis, Professor für Digital Health: Human-Centered Transformative AI am Hasso-Plattner-Institut: "Erstens braucht man Zugriff auf das KI-Modell, das angegriffen werden soll, etwa über das Netzwerk einer Klinik. Zweitens braucht man Zugriff auf einen Datenpunkt, von dem man wissen will, ob er ins Modell

eingeflossen ist. Vorstellbar wäre beispielsweise, dass solche Datenpunkte bei einem Hackerangriff erbeutet werden. Drittens braucht man eigene KI-Infrastruktur - Rechner, auf denen KI-Modelle laufen, die auf dem gleichen Datentyp basieren wie das attackierte Modell."

Mit diesem Setup ließe sich beispielsweise ein KI-Modell angreifen, das aus Blutbildern die Erfolgsaussichten für eine Krebs-Immuntherapie ableitet. Für sich genommen gibt das Blutbild keine Auskunft über eine Erkrankung. Kann ein Angreifer aber zeigen, dass ein Datenpunkt in das Modell eingeflossen ist, lässt sich mit größerer Wahrscheinlichkeit sagen, dass die Patientin oder der Patient an Krebs erkrankt ist oder war.

Mögliche Folgen für Betroffene

Solche digitalen Attacken können schwerwiegende reale Folgen haben, das macht Moritz Knolle an einem fiktiven Beispiel deutlich: "Stellen Sie sich vor, Sie wurden wegen einer Krebserkrankung behandelt und haben Ihre Daten für die Forschung zur Verfügung gestellt", sagt der Medizininformatiker. "Jahre später - der Krebs ist seitdem nicht wieder aufgetreten - wollen Sie eine private Zusatzversicherung abschließen. Nun hat aber ein Angreifer herausgefunden, dass Ihre Daten für das Training eines Modells zur Tumoranalyse verwendet wurden. Diese Information gelangt an den Versicherer, zum Beispiel über Datenanalysen von Drittanbietern oder entsprechende Risikoprofile. Sie werden daraufhin als Hochrisikopatient mit den entsprechenden Beiträgen eingestuft - und erfahren unter Umständen nie den Grund dafür."

Die MIAs waren besonders häufig erfolgreich, wenn attackierte Individuen zu einer Gruppe gehörten, die in dem Datenset nur wenig vertreten war. Das können bestimmte Merkmale der Organe in der Bildgebung sein, aber auch Daten von Minderheiten. "Das ist besonders schwerwiegend, weil Diskriminierung durch KI auch in der Medizin eine Rolle spielt und beispielsweise manche Modelle weniger genauere Vorhersagen treffen, wenn die Patientin oder der Patient einer Minderheit angehört", sagt Daniel Rückert.

Leistungsstarke Modelle besonders betroffen

Die Forschenden konnten zeigen, dass die Attacken umso erfolgreicher waren, je größer und komplexer das Modell war. Dass gerade leistungsfähige Modelle angreifbar sind, zeigt aus Sicht der Forschenden, dass sich die Problematik in den kommenden Jahren ohne Gegenmaßnahmen deutlich verschärfen könnte.

Sie sprechen sich daher dafür aus, die Risiken neuer Modelle vor deren Veröffentlichung künftig immer auf Ebene der Individuen zu überprüfen. Weitere Gegenmaßnahmen sind eine strenge Kontrolle des Zugangs zu KI-Modellen. "Es gibt außerdem bereits heute effektive Schutzmaßnahmen gegen MIAs, die direkt beim Trainieren der Modelle zum Einsatz kommen können. Beispielsweise werden bei der 'Differential Privacy' kleine Veränderungen in die Trainingsdaten eingebaut, die die Berechnungen des Modells nicht stören, aber eine MIA deutlich erschweren", sagt Georg Kaissis.

Publikation:

Knolle, M.A., Menten, M.J., Jungmann, F. et al. Disparate privacy risks from medical AI. Nature (2026).
https://doi.org/10.1038/s41586-026-10688-0