MedWiss.Online

Hautkrebsdiagnosik: KI-Modelle unter Alltagsbedingungen

Datum: 22.06.2026

Original Titel:
Limits of Artificial Intelligence Models for Skin Cancer Diagnosis in Realistic Settings.

Kurz & fundiert

- Künstliche Intelligenz (KI-Modelle) zur Hautkrebsdiagnose: Relevant unter Bedingungen des klinischen Alltags?
- Vergleich KI versus ärztliche Expertise
- Ärzte unterschiedlicher Institutionen mit Berufserfahrung zwischen 1 und ≥10 Jahren
- 1 117 Fälle mit klinischen und dermatoskopischen Bildern
- KI übertrifft Ärzte mit < 3 Jahren Erfahrung
- Ärzte mit > 10 Jahren Erfahrung besser als KI-Modelle

MedWiss – **Unter standardisierten Bedingungen erreichen Modelle künstlicher Intelligenz (KI) bereits eine hohe Genauigkeit bei der Diagnose bösartiger Hauterkrankungen. Die Patientenversorgung geht aber über die reine Interpretation von Bildern hinaus: Auch seltene und atypische Erscheinungsbilder gehören zum klinischen Alltag. In diesem Setting können ärztliche Fähigkeiten, wie eine aktuelle Studie zeigt, die diagnostische Performance aktueller KI-Modelle übertreffen.**

Künstliche Intelligenz (KI) hat bei der Interpretation von Bildern in der medizinischen Diagnostik deutliche Fortschritte erzielt und in manchen Bereichen damit die ärztliche Expertise erreicht oder sogar übertroffen. Allerdings kamen die KI-Modelle meist unter kontrollierten Studienbedingungen und mit eingegrenzten Aufgaben zum Einsatz. Welche Genauigkeit eine KI bei komplexeren Aufgaben erreicht, beispielsweise bei ärztlichen Aufgaben in der Hautkrebsdiagnose, untersuchte eine französische Arbeitsgruppe.

Diagnostische Genauigkeit: Wie gut ist KI unter Alltagsbedingungen?

Die Studie vergleich die Leistung von Hautärzten mit der von KI-Modellen. Die teilnehmenden Ärzte (Alter 29 – 37 Jahre) werteten jeweils 100 zufällig ausgewählte Bilder aus einem Datenset mit 1 117 Fällen aus. Der größte Teil der Ärzte (73,3 %) hatte weniger als 3 Jahre Dermatoskopie-Erfahrung. Im Vergleich zu den Ärzten wurden drei KI-Modelle eingesetzt: Ein neuronales Faltungsnetz (convolutional neural network, CNN; ResNet50) sowie zwei KI-Modelle, die auf große Datensätze trainiert wurden (Foundation Models: PanDerm uni- und multimodal). Als primäre Zielgröße wurde die Genauigkeit der Diagnosen erfasst. Die sekundären Zielgrößen umfassten Sensitivität der Einstufung als gut- oder bösartig, Spezifität und ausgewogene Genauigkeit (balanced accuracy), mit

der die Leistung eines Klassifizierungsmodells bewertet wird.

Vergleich von KI-Modellen mit Hautärzten in der Hautkrebs-Diagnostik

Die mittlere Genauigkeit lag für CNN bei 56,7 %, für das unimodale Modell bei 72,2 % und für das multimodale Modell bei 66,3 %. Die Genauigkeit der ärztlichen Diagnosen lag für Ärzte mit einem Erfahrungshorizont < 1 Jahr bei 59,1 % bis hin zu 74,2 % für diejenigen mit > 10 Jahren Erfahrung. Im Durchschnitt betrug die Genauigkeit der ärztlichen Diagnosen 65,9 %.
Die ärztliche Diagnostik erwies sich gegenüber dem CNN-Modell als überlegen (65,9 % vs. 56,7 %; Differenz 9,2 %; 95 % Konfidenzintervall, KI: -9,8 – 8,5 %; $p < 0,001$). Die Genauigkeit des zweiten KI-Modells (unimodal) übertraf die von Ärzten mit weniger als 3 Jahren Erfahrung (72,2 % vs. 68,2 %; Differenz: 4,0 %; 95 % KI: 3,2 – 4,9 %; $p < 0,001$).

Lange ärztliche Erfahrung ist KI-Modellen überlegen

Die Autoren schließen, dass Experten mit mehr als 10 Jahren Erfahrung in der Hautkrebsdiagnostik die Genauigkeit moderner KI-Modelle übertrafen. Spezifische KI-Modelle (Foundation Models) übertrafen aber die Genauigkeit der Diagnosen von Ärzten mit weniger als 3 Jahren Erfahrung und lagen gleichauf mit Ärzten mit 3 – 10 Jahren Erfahrung.

Referenzen:
Anriot J, Yan S, Coste C, Tschandl P, Verlingue L, Andremasse C, Amini-Adle M, Perrot JL, Ge Z, Kittler H, Thomas L. Limits of Artificial Intelligence Models for Skin Cancer Diagnosis in Realistic Settings. JAMA Dermatol. 2026 Jun 3:e261492. doi: 10.1001/jamadermatol.2026.1492. Epub ahead of print. PMID: 42234423; PMCID: PMC13234745.