

# Diagnostische Leistung und Prompt-Injection Anfälligkeit multimodaler KI in der dermatoskopischen Melanomdiagnostik

**U. Cambaz<sup>1</sup>**, I. Güler<sup>2</sup>

<sup>1</sup> Eberhard-Karls-Universität Tübingen, Medizinische Fakultät, Tübingen, Baden-Württemberg, Deutschland

<sup>2</sup> Otto-von-Guericke-Universität Magdeburg, Universitätsklinik für Plastische, Ästhetische und Handchirurgie, Magdeburg, Sachsen-Anhalt, Deutschland

## Inhalt

### Einleitung

Multimodale KI-Systeme werden zur Beurteilung dermatoskopischer Hautläsionen eingesetzt. Neben ihrer diagnostischen Leistung rückt ihre Robustheit gegenüber Manipulationen der Eingabedaten in den Fokus. Ziel dieser Studie war die Evaluation der diagnostischen Leistung und der Anfälligkeit der Vision-Language-Modelle gegenüber Prompt-Injection.

### Methoden

52 histopathologisch gesicherte dermatoskopische Bilder (26 invasive Melanome, 26 melanozytäre Nävi) wurden mit Claude Opus 4.7, Gemini 3.1 Pro und GPT-5.4 analysiert. Untersucht wurden drei Prompt-Injection-Szenarien (visuell, Metadaten, kombiniert). Primärer Endpunkt war die diagnostische Genauigkeit.

### Ergebnisse

Unter Standardbedingungen erreichten die Modelle eine diagnostische Genauigkeit von 58–62%. Bereits ein einzelnes irreführendes Wort führte in allen Angriffsszenarien zu einem nahezu vollständigen Verlust der diagnostischen Genauigkeit. Die Antworten blieben auch bei wiederholten Analysen nahezu unverändert, ebenso die angegebene Sicherheit. Die diagnostische Entscheidung ließ sich gezielt beeinflussen.

### Schlussfolgerung

Aktuelle multimodale KI-Systeme weisen in der dermatoskopischen Melanomdiagnostik Limitationen der diagnostischen Leistung und eine ausgeprägte Anfälligkeit gegenüber Prompt-Injection auf. Für einen sicheren klinischen Einsatz sollten neben der diagnostischen Genauigkeit auch Robustheit gegenüber Manipulationen, technische Schutzmechanismen und ärztliche Validierung berücksichtigt werden.