Expert-Informed Melanoma Classification with Dermoscopic and Histopathological Data: Results and Recommendations for Practice

Haggenmüller S[1], Heinlein L[1], *Abels J [1], Winterstein JT[1], Wies Ch[1], Kuehn A[1], Mehrtens H[1], Utikal JS, Wohlfeil SA, Meier F, Hobelsberger S, Gellrich FF, Sergon M, Hauschild A, French LE, Heinzerling L, Schlager JG, Ghoreschi K, Schlaak M, Hilke FJ, Poch G, Korsing S, Berking C, Hepp MV, Erdmann M, Haferkamp S, Drexler K, Schadendorf D, Sondermann W, Goebeler M, Schilling B, Kather JN, Fröhling S, Llamas-Velasco M, Requena LC, Ferrara G, Fernandez-Figueras M, Fraitag S, Müller CS, Starz H, Kutzner H, Barnhill R, Carr R, Braun SA, Holland-Letz T, Brinker TJ[1]*

GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer

MOTIVATION

Deep learing models have achieved human-level performance in **skin cancer classification tasks.** However, uncertainties in ground truth labeling – particularly in borderline cases – can compromise model reliability and clinical safety.

To adress this, we developed expert-informed classifiers for **dermoscopic** and **histopathological whole slide images**, trained on **majority-labels** and **soft-labels** (i.e., incorporating reviewer uncertainties) annotated by an **international expert panel** of eight (dermato-)pathology experts.

METHODS

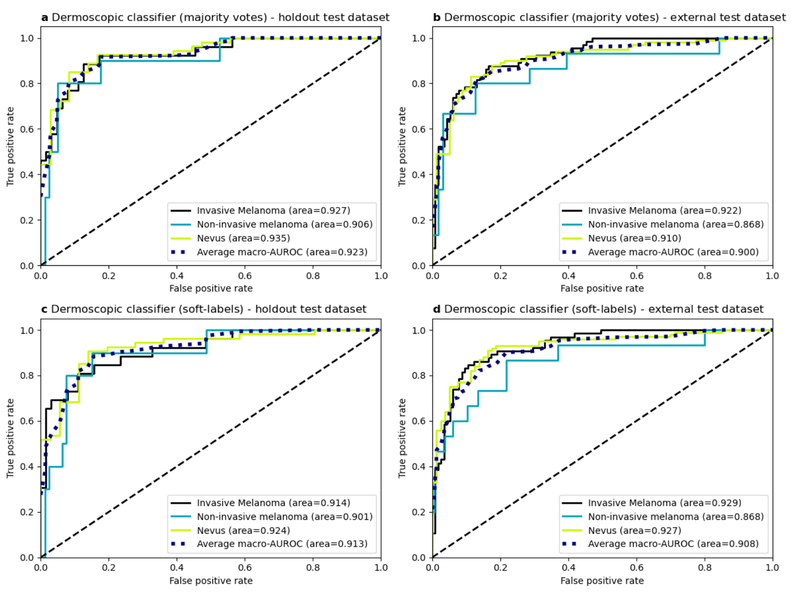
Models were evaluated on a **holdout** and an **external test dataset** comprising 792 **prospectively collected**, clinically representative cases from eight German university hospitals.

Data were acquired using **standardized HEINE dermatoscope setups** and distributed across **three classes**: invasive melanomas, non-invasive melanomas, and nevi.

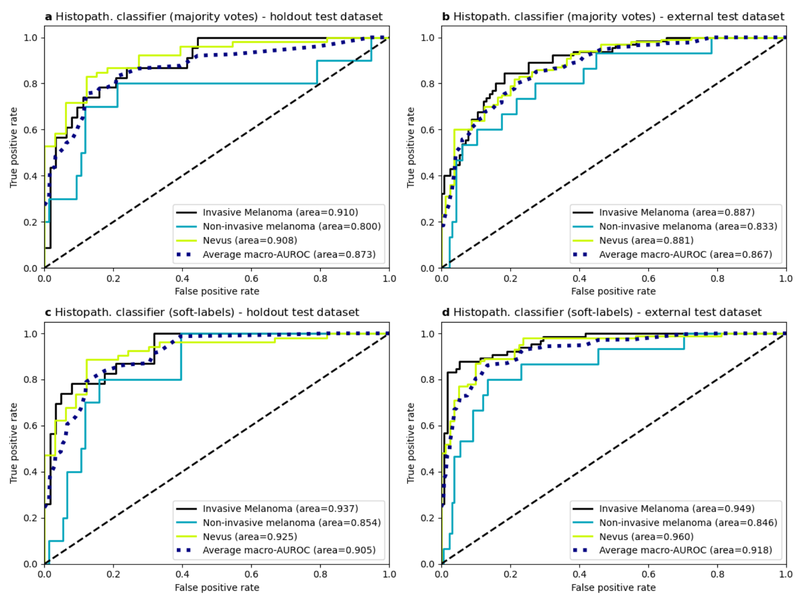
Furthermore, the algorithms, inherently embedding expert knowledge, were compared to **local dermatologists** and **pathologists** to assess their clinical utility.

RESULTS and DISCUSSION

The **dermoscopic classifiers** achieved AUROC scores of 0.923 (0.866-0.969) and 0.900 (0.845-947) using **majority votes**, and 0.913 (0.855-0.962) and 0.908 (0.858-0952) using **soft-labels**.



The **histopathological classifiers** achieved AUROC scores of 0.873 (0.787-0.941) and 0.867 (0.807-0.916) using **majority votes**, and 0.905 (0.851-0.952) and 0.918 (0.874-0.954) using **soft labels**.



Compared with **local dermatologists** and **pathologists**, expert-informed models demonstrated comparable performance while exhibiting **complementary diagnostic strengths**, supporting their use as **digital second-opinion tools** in routine clinical care.

By extending beyond binary classification to a **multiclass setting** and **incorporating expert uncertainty** through soft-label training, our approach better reflects **real-world diagnostic complexity** without compromising performance.

Notably, both human experts and AI systems encountered difficulties with **non-invasive melanomas**, underscoring their **intrinsic diagnostic ambiguity**.

These findings highlight the **importance of uncertainty-aware**, human-in-the-loop AI systems and emphasize the need **for high-quality prospective datasets** and ground truth definitions beyond conventional histopathology alone to enhance the clinical **robustness, safety**, and **translational relevance** of diagnostic models.

Dr. Sarah Haggenmüller
AI for skin cancer diagnostics

@DRSARAHHAGGENMUELLER

[1] Division of Digital Prevention, Diagnostics and Therapy Guidance, German Cancer Research Center (DKFZ), INF 223, 69120 Heidelberg, Germany.
Contact: sarah.haggenmueller@dkfz.de and titus.brinker@dkfz.de