Supplementary Material

S1. Background-swap experiment

A recurring concern with the Royal Prince Alfred Hospital subset is that its uniform white background acts as a shortcut. To probe this, we took the 16 images from that source and replaced the white background with black, then repeated the experiment with a second variant in which the surgical skin markers were also painted out. Table 7 reports the mean predictive entropy on the original images and the change (Δ) after each edit, for both the binary head (H1) and the five-class head (H2).

Table 7: Mean predictive entropy on the 16 Prince Alfred images and its change after editing the background. Lower entropy means the model became more confident. "Black bg" replaces the white background with black. "+markers removed" additionally erases the surgical markers.

	Original		Black bg (Δ)		+markers removed (Δ)	
Method	H1	H2	H1	H2	H1	H2
Deep Ensembles	0.504	0.475	−0.074	+0.068	−0.114	+0.075
MC Dropout	0.224	0.234	−0.066	−0.045	−0.097	−0.099
DropConnect	0.283	0.286	−0.219	−0.188	−0.158	−0.154
Flipout	0.590	0.805	−0.119	−0.106	−0.135	−0.099
DUQ	0.408	0.462	−0.092	−0.052	−0.174	−0.052

With only $n = 16$ images the result is suggestive rather than conclusive. Entropy on H1 falls for all five methods once the white background is gone, and H2 entropy falls for four of the five. Deep Ensembles is the exception: its H2 entropy rises slightly in both variants. Removing the markers on top of the background swap deepens the drop for most methods but does not change the direction. We do not report accuracy deltas, because at $n = 16$ each flipped case moves accuracy by about 6.3 points, so the numbers are too noisy to interpret. The pattern fits a source-specific background artifact, yet because the edit also changes lighting and contrast around the lesion, it is not a clean isolation of the background alone.

S2. Consolidated diagnosis mapping

The dataset records 22 fine-grained secondary diagnosis categories, which we collapse into five classes for the H2 task. The groupings are:
Benign. Benign melanocytic proliferations, benign epidermal proliferations, benign (other), benign soft tissue proliferations (fibro-histiocytic, vascular, neural), benign adnexal epithelial proliferations (sebaceous, follicular, apocrine or eccrine), cysts, hemorrhagic lesions, and mast cell proliferations.
Melanoma. Malignant melanocytic proliferations (melanoma).
Malignant Epidermal. Malignant epidermal proliferations.

Malignant Non-Epidermal. Malignant adnexal epithelial proliferations (follicular, apocrine or eccrine) and malignant soft tissue proliferations (fibro-histiocytic, vascular).
Other. Indeterminate epidermal and melanocytic proliferations, flat melanotic pigmentations that are not melanocytic nevi, inflammatory or infectious diseases, collisions (benign-only, or with at least one malignant component), and exogenous.

S3. Qualitative examples

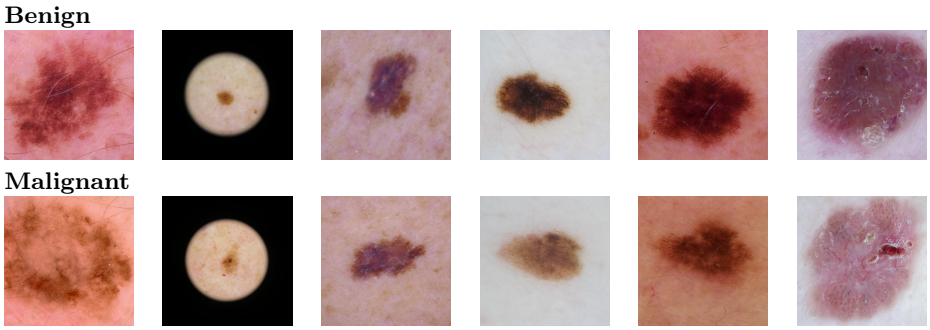

Fig. 4: Examples of visually similar benign–malignant lesion pairs from the ISIC archive. Each column shows a benign (top) and malignant (bottom) lesion paired by similarity in a neural network's learned feature representation. The visual diversity across columns demonstrates intra-class variation.

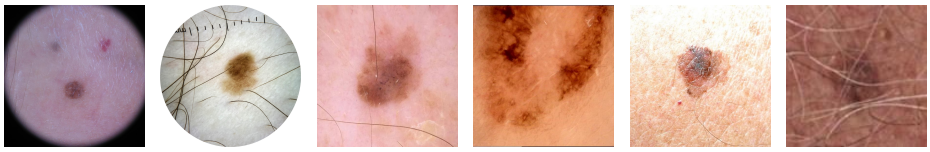

Fig. 5: Six melanoma images from different data sources and imaging modalities. The images differ substantially in background color, magnification, lighting, and resolution due to differences in clinical equipment and imaging protocols.

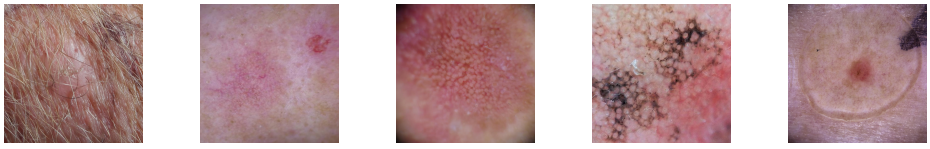

Fig. 6: Five indeterminate-set samples with the highest predictive entropy on the multiclass head, ranked by Deep Ensembles.