Ensemble uncertainty estimation improves skin cancer malignancy prediction

W. Max Schreyer BA[1, 2, 4], Ravi Samatham PhD[3], Elizabeth Berry MD[3] and Reid F. Thompson MD, PhD[1, 2, 4, 5, †]

[1] Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR, United States
[2] Knight Cancer Institute, Oregon Health and Science University, Portland, OR, United States
[3] Department of Dermatology, Oregon Health and Science University, Portland, OR, United States
[4] Portland VA Research Foundation, Portland, Oregon, USA
[5] Department of Radiation Medicine, Oregon Health & Science University, Portland, Oregon, USA
[†] Corresponding author: thompsre@ohsu.edu

Abstract

Widespread access to imaging technologies and stronger machine learning (ML) architectures for dermatology tasks such as malignancy prediction have spurred a race to develop models to assist in the automated diagnosis of skin cancer. However, high diagnostic performance on benchmarking datasets quickly deteriorates when models are challenged with data from disparate clinical sources. Generalization gaps stem from the high variability in skin lesion images due to lighting, capture angle, imaging technology and patient phenotype among other factors, impeding the safe application of diagnostic ML models in practice. In this study, we apply a novel ensemble uncertainty-estimation approach to detect out-of-distribution skin lesion images from four publicly available datasets across five countries. Using our method, Supervised Autoencoders for Generalization Estimates (SAGE), we quantify likeness of images from patients in Argentina, Brazil and the United States to the popular HAM10000 benchmarking dataset and identify problematic image artifacts that affect the reliability of predictions in a pre-clinical setting. We show how filtering images based on SAGE score thresholds can improve the performance of a separate malignancy prediction model and how our approach is robust to variations in image modality and the introduction of new diagnostic classes, providing users with a powerful tool for interrogating key differences between their data and the training distribution of an ML model before clinical implementation.

Introduction

Skin cancers are the most commonly-diagnosed malignancy worldwide, with over 300,000 cases of melanoma (age-standardized incidence rate (ASIR) = 3.56 per 100,000 people), 1.9 million cases of squamous cell carcinoma (ASIR = 22.38) and 4.4 million cases of basal cell carcinoma (ASIR = 74.10) occurring globally in 2021.[1,2] However, malignancies represent only a fraction (22%) of all index skin lesions referred to dermatologists for biopsy and review, resulting in a clinically significant tension between diagnostic sensitivity and specificity.[3] Compared to dermatologists, Primary Care Providers (PCPs) receive significantly less training in dermatologic conditions resulting in decreased diagnostic ability. Gaps in clinicians practicing dermatology also exist and vary by level of training, years of experience, and are further amplified by use of dermoscopy.[4] Further barriers to skin cancer detection combine to exacerbate pre-existing health disparities such as patients with low socioeconomic status also living in rural locations.[2,5] Fortunately, the combination of widely available imaging technologies and automated detection through machine learning (ML) have the potential to revolutionize skin cancer diagnosis and improve global health equity as a result.[6,7]

Many ML models have been trained to identify cutaneous malignancies within skin imaging datasets.[8–16] However, there are notable differences between highly-curated benchmarking datasets used to train and evaluate ML models and the images encountered in real world clinical settings, often leading to reduced model effectiveness in practice.[4,17,18] Even state-of-the-art models suffer from these performance drops; the top 25 entries from the ISIC 2019 Grand Challenge misclassified nearly 50% of previously-unseen skin lesion images despite stellar performance on benchmarking datasets.[19] Ambitious ML-assisted dermatology smartphone phone apps have also been shown to have low accuracy and can therefore pose a harm to users attempting to self-diagnose potentially life-threatening diseases without expert oversight.[20–22] One major source of difficulty in automating skin cancer diagnosis is the variability in imaging technologies like dermoscopes which cannot properly standardize inputs to predictive algorithms – outputs can vary based on whether or not the dermatoscope is polarized and the type of light source.[23,24] Indeed, even the presence of modest artifacts such as blur or blue/red shifted-pixel intensities showed marked decreases to ML performance for both diagnostic and disease management tasks when compared to control images.[25] Beyond device or light-induced inconsistencies, patient phenotypes such as dark skin have been shown to degrade ML model accuracy and represent an ethical dilemma in the deployment of algorithms trained on benchmarked datasets which either omit skin type metadata or severely underrepresent some patient populations.[26,27] It is therefore critically important to assess the similarity of new dermatological images with a model's training data on a case-by-case basis, in order to reduce disparities and identify potentially problematic samples before a model is used in any treatment pipeline.

Some diagnostic dermatology models have attempted to remove corrupted or low-quality samples through out-of-distribution (OOD) detection or by building inherent measures of uncertainty into the prediction task.[28–30] This typically combines uncertainty estimation with another primary task, lacking flexibility in adapting to different tasks or pairing with stronger models as they become available. Other single-pronged approaches to measuring uncertainty

such as calibration of classifier softmax outputs can be intrinsically measured at time of prediction, but fail silently under scenarios of data drift.[31] Our previous work has introduced an ensemble approach (Supervised Autoencoder for Generalization Estimates [SAGE]) for image uncertainty estimation that is robust to corruptions and perturbations missed by singular OOD metrics, and can be paired with any downstream model's training data.[32] In this study we implement SAGE to assess the generalization potential of skin cancer detection algorithms across dermatology images from five countries and multiple imaging modalities. We use our SAGE scoring system to improve performance of a pre-trained malignancy predictor, enriching for images that are appropriate versus problematic for ML-assisted diagnosis.[26]

Results

We obtained four publicly-available skin lesion imaging datasets (HAM10000[33], HIBA[34], UFES[35] and DDI[26]) each with distinct characteristics and metadata (Supplementary Table 1; See Methods). HAM10000, the most widely used dataset for benchmarking, contains only one image size ([600 x 450] pixels) whereas images from other datasets vary in resolution and dimensions (range: [147 x 147] - [4,128 x 3,176] pixels). HAM10000 features lesions with seven diagnostic classes, five considered benign (actinic keratosis, benign keratosis, dermatofibroma, melanocytic nevi, and vascular skin lesion) and two considered malignant (basal cell carcinoma, melanoma). (Supplementary Fig. 1) Notably absent from the HAM10000 dataset are examples of squamous cell carcinoma, a common form of keratinocytic malignancy that comprises a significant minority of the images present in the HIBA (9.78%) and UFES (8.12%) datasets. Even more substantially, nearly half of all lesions in the DDI dataset (45.12%) represented 36 other diagnoses not present in the HAM10000 dataset. (Supplementary Table 2) Both dermatoscope and clinical smartphone imaging modalities are included in the HIBA and UFES datasets where the ratio of dermoscopic to smartphone images is 3.67:1 for HIBA and not provided for UFES.

Using HAM10000 as a primary reference dataset, we trained a SAGE model to simultaneously encode an image into a 32-dimensional compressed latent space vector, reconstruct the original image and predict its diagnostic class from the compressed embedding. (Fig. 1) We found that the SAGE classifier module performed well on HAM10000 holdout test images (weighted F1 score = 0.842) while the decoder was able to coarsely reconstruct lesion size, shape and pigmentation color (Supplementary Fig. 2). Additionally, the SAGE encoder latent space could distinguish variation within and between different image sets (e.g. HAM10000, training and testing distributions are highly similar whereas HIBA, UFES, and DDI occupy areas of training data paucity) (Supplementary Fig. 3). To quantify conformity between the training and test images, we calculated each image's SAGE score – an ensemble metric for assessing similarity based on the reference distributions for the model's latent embedding distance, classifier confidence and reconstruction error (Methods). The distributions of SAGE scores for HAM10000 testing and non-HAM10000 datasets (i.e. HIBA, UFES, DDI) were significantly lower than the train score distribution, with the HAM10000 test images showing the highest median score (median = 0.50) and DDI the lowest (median = 0.06). (Fig. 2A) These dataset-level differences were largely consistent across diagnoses. (Fig. 2B) However, we observed that images associated with diagnoses missing from the HAM10000 dataset had a lower average SAGE

score than those corresponding to diagnoses present in the training data (mean = 0.12 vs. mean = 0.22). When we removed images of lesions with diagnoses not present in the HAM10000 dataset from HIBA, UFES and DDI, we expected that differences in SAGE score might be clearly delineated by metadata such as imaging modality and Fitzpatrick Skin Type (FST)[36]. However, we found a surprising degree of overlap in these distributions (Dermoscopic, IQR: [0.09 - 0.32]; Smartphone, IQR: [0.04 - 0.16]; FST I-II, IQR: [0.07 - 0.26]; FST III-IV, IQR: [0.07 - 0.24]; FST V-VI, IQR: [0.03 - 0.14]) pointing to the presence of other distinguishing image attributes that were unidentifiable from the metadata or lesion type alone. (Fig. 2C-D)

We note that HIBA, UFES and especially the DDI dataset are laden with low SAGE scoring images with many out-of-distribution features not present in the HAM10000 data, such as measuring devices (e.g. rulers), dense hair, skin coverings and other markings. (Fig. 3C, D, E) Even within the HAM10000 training and test data we identify uncommon edge cases with low SAGE scores such as lesions cropped by the image frame or bubbles resulting from a lapsed contact between the dermoscope lens and skin. (Fig. 3A, B) To establish how the presence and severity of out-of-distribution features affect SAGE score, we manually-annotated 12 independent features for all HIBA, UFES and DDI images (n = 4,527). (Supplementary Fig. 4, Supplementary Table 3) Some features had clear negative associations with SAGE scoring such as the presence of a ruler or use of camera flash which showed 51.4% and 10.7% reductions, respectively. Similarly, the presence of non-skin background in an image had a severe negative effect on SAGE score even at low (44.4%) and medium (74.5%) levels. Linking multiple features together with a SAGE score overlay gives a concise description of what constitutes a typical “high-quality” image, such as having no or low amount of hair, high contrast between the lesion and skin and little to no non-skin background. (Fig. 4A) To that end, we frequently observed compounding negative effects when detrimental image features were stacked such as the 94.7% reduction to SAGE score when both dense hair and non-skin background are co-present (mean = 0.01) vs. a 63.2% decrease for dense hair only (mean = 0.07) and a 68.4% decrease for majority non-skin background alone (mean = 0.06). (Fig. 4B) We also noted that the presence of some image features did not affect patients with different skin phototypes equally. For instance, inclusion of a ruler had a 38.9% reduction to average SAGE score in patients with lighter skin phototypes (FST I-II) whereas this decreased by 75.0% for patients with dark skin (FST V-VI), possibly resulting from the higher contrast between the illuminated measuring devices and darker skin. (Fig. 4C) Intriguingly, statistically significant difference between SAGE score distributions for the lowest (I-II) and highest (V-VI) and FST levels for were ablated after low-quality images were removed (Wilcoxon rank sum test, p=0.06), possibly illustrating how extraneous image features are a larger detriment to image quality than changes to some underlying patient phenotypes. (Supplementary Fig. 5)

We next evaluated the ability of SAGE score to improve model performance when deployed as a prospective filter for OOD images. Using a mild SAGE score threshold (0.2), we observed that HAM10000 images were retained at high rates (train = 83.9%, test = 87.2%), whereas the majority of images from other datasets were identified as outliers (e.g. DDI retained only 8.4% of its data). (Fig. 5A-B) This selective filtering demonstrated significant improvements to overall performance of a pre-trained open source predictor[26] as assessed by area-under the ROC

curve (AUC) in both the training (original AUC = 0.96, filtered AUC = 0.98) and mixed test datasets (original AUC = 0.82, filtered AUC = 0.95), with progressively larger performance improvements as SAGE threshold increased (Fig. 5C-D). Positive predictivity of the mixed test set also improved to equal the average precision (AP) performance of the training data at the highest threshold level tested (AP = 0.87, threshold = 0.4) (Fig. 5E-F). In contrast to SAGE score thresholding, we found that filtering images based on a series of increasing model-intrinsic confidence scores did not improve performance on a mixed test dataset and resulted in a consistent decline in the percentage of true malignant samples (Supplementary Fig. 6). Importantly, we found that SAGE score filtering improved malignancy prediction performance for both smartphone and dermoscopic images and all skin types, with patients at the highest FST levels (V-VI) seeing the largest improvements to model performance (Supplementary Fig. 7).

Using SAGE, we find that images of previously-unseen lesion types such as squamous cell carcinoma, T-cell lymphomas, kaposi sarcoma and metastatic carcinoma have substantially lower overall score distributions than e.g. melanoma and basal cell carcinoma, demonstrating how new disease classes can be isolated by their dissimilarity to the training data. (Supplementary Fig. 8A) Additionally, we discovered that new disease images had poor zero-shot performance when evaluated on the pre-trained malignancy prediction model, and that performance for these classes was poor across all SAGE score thresholds – accuracy for squamous cell carcinoma, t-Cell lymphomas, kaposi sarcoma and metastatic carcinoma declined by 3.6% whereas melanoma and basal cell carcinoma accuracy increased by 19.6% after thresholding at a SAGE score of 0.2. (Supplementary Fig. 8B, C) For each disease class, we also identified false negative (FN) instances where malignant lesions fell below the predictor's decision boundary (0.733) and analyzed the spread of points above and below a conservative SAGE score cutoff (0.151) where 90% of the training data is retained. (Supplementary Fig. 9) This revealed an enrichment of FN images falling below the score cutoff for all malignant classes, with previously-unseen malignancies showing a high proportional capture of FN examples (mean = 0.88).

Discussion

In this study we demonstrate the successful application of our novel uncertainty estimation method to enhance the reliability of downstream skin cancer malignancy detection. Our work uses the most popular benchmarking dataset of dermoscopic lesion images, HAM10000, as a reference to quantify out-of-distribution test images and identify detrimental image attributes without selection bias or requiring a classification label. To our knowledge, this is the first study to apply ensemble uncertainty estimation to dermatology imaging datasets in a model-agnostic manner. By incorporating four datasets from five countries in our analysis, we also address the critical need to reduce generalization gaps when applying skin cancer prediction models across global populations. Our method can pair with any dermatology ML task so long as the training data is known, which will improve integration of robust uncertainty estimation into more trustworthy automated pipelines. We envision SAGE as a powerful extension to or replacement for the concept of model cards, where users can view detailed information about a model's training data.[37] In our case, we not only quantify the differences between training data and

prospective test data but also encourage users to interactively probe examples to see where and how they might differ from distributions of training image features.

However, this analysis has several limitations. Despite sourcing images from four independent datasets, the vast majority of images pertain to patients with lighter skin phototypes (FST I-IV) and lack significant representation from Asian and African populations. We also note that the FST scale itself has been shown to vary based on environmental conditions.[38,39] The data in this study consists of a relatively small cohort of ~15,000 training and test examples; inclusion of larger imaging datasets or pre-training on dermatological images could strengthen encoder fine-tuning and yield more relevant embeddings instead of initializing on default ImageNet weights from PyTorch. We are also unable to assess the influence of identifiable markings such as scars or tattoos, as these were explicitly removed by the authors of each benchmarking dataset for patient de-identification purposes. We did not manually-annotate features for HAM10000 images due to the large dataset size and the incidence of these features in the SAGE training data was not analyzed. Furthermore, other image features that have been shown to affect ML model performance on skin imaging datasets such as color balance (e.g. blue or red pixel intensity shifts) were not annotated or assessed. Additionally, we did not study the effects of anatomical location on SAGE scoring which could provide additional guidance to users of the downstream malignancy prediction model.

While our model architecture and encoder embedding size of 32 dimensions was selected to lower overall training loss, we did not explore larger latent spaces >100 dimensions or other model architectures which might facilitate improved image reconstructions and better classification performance. As hardware support for training large ML models continues to decrease in cost, our work could expand to encompass larger foundational encoders and more complex latent space embeddings. Future work will also seek to pair SAGE with more difficult image segmentation, skin cancer diagnosis and lesion monitoring tasks across patient groups and will test a larger array of downstream models to develop an interval of task improvement facilitated by SAGE. Finally, while this work was explicitly focused on skin cancer detection using dermoscopy and clinical photography, SAGE could be broadly applicable for a wide array of clinical and non-clinical use cases.

Methods

Datasets

Humans Against Machine (HAM10000)[33] consists of 10,015 dermoscopy photos sourced from the Rosendahl dermatological practice in Australia and the ViDIR Group in Austria. We downloaded files from the Harvard Dataverse (https://doi.org/10.7910/DVN/DBW86T) containing JPEG images and associated metadata as a CSV file. The metadata file contains columns detailing lesion and image identifiers, patient age at time of image capture (median = 50 years old), body site of the lesion and diagnostic method. All images are of pigmented lesions and ground-truth labels are included in the metadata with six diagnostic classes: actinic keratosis (*ak*), basal cell carcinoma (*bcc*), benign keratosis (*bk*), melanocytic nevus (*nevi*), melanoma

(*mel*) and vascular skin lesion (*vasc*). For the purposes of this study, only *bcc* and *mel* classes were considered malignant; all other classes were considered benign, although *ak* is known to progress to cancer in certain cases. All malignant lesions were histologically-confirmed. Images have an original size of 600 x 450 pixels and were resized using bilinear interpolation where the smaller dimension of height or width was resized to 299 pixels, preserving aspect ratio. Images were cropped to 299 x 299 pixels after resizing to ensure consistency across datasets and compatibility with open-source ML architectures. Authors include multiple images of the same lesion with differing perspectives as a form of natural data augmentation and perform quality control to remove out-of-focus images and images with insufficient zoom of small lesions. We randomly-split HAM10000 into training and test sets stratified by diagnostic class and removed any images of lesions from the test set that were also present in the training set, leaving 9,013 images for training and 600 for testing.

The **Hospital Italiano de Buenos Aires (HIBA)**[34] dataset contains 1,616 JPEG images of mixed dermoscopic and clinical smartphone images from Argentina which were downloaded from the ISIC archive (https://doi.org/10.34970/587329). The metadata file contains a row for each image which includes a unique image and lesion ID as well as an anonymized patient identifier. Separate metadata columns are also included for age (median = 65 years old), sex, family history of skin cancer, Fitzpatrick Skin Type (FST) and image type differentiating between clinical smartphone and dermoscopic imaging technologies. 93% (*n*=566) of the 623 patients had skin type data and age and sex were recorded for over 99% of patients. Each image is assigned a ground-truth label from one of 10 diagnostic classes, 9 of which overlap with HAM10000 categories with the exception of squamous cell carcinoma (*scc*), an additional malignancy not included in the training data. The remaining image classes were basal cell carcinoma (*bcc*), melanoma (*mel*), melanocytic nevus (*nevi*), actinic keratosis (*ak*), dermatofibroma (*df*), vascular skin lesion (*vasc*), solar lentigo (*bk*), seborrheic keratosis (*bk*) and lichen planus-like keratosis (*bk*). All malignancies for *bcc, scc* and *mel* lesions were biopsy-confirmed. Images from the HIBA dataset vary in size between a maximum of 4,128 x 3,096 pixels and a minimum of 162 x 152 pixels. All files were preprocessed by converting to 8-bit RGB color channels, resized with bilinear interpolation and center-cropped to 299 x 299 pixels before use in our study.

The dermatology program at the **Universidade Federal do Espírito Santo (UFES)**[35] in Brazil published a dataset in 2020 containing 2,298 dermoscopic and clinical smartphone images with 6 diagnostic classes: basal cell carcinoma (*bcc*), squamous cell carcinoma (*scc*), actinic keratosis (*ak*), melanoma (*mel*), melanocytic nevus (*nevi*) and seborrheic keratosis (*bk*). All *bcc, scc* and *mel* images were considered malignant and lesions were biopsy-confirmed. We downloaded images in PNG format and the metadata CSV file from the Mendeley Data Commons link provided by the authors (https://doi.org/10.17632/zr7vgbcyr2.1). The metadata file contains patient, lesion and image identifiers as well as detailed lifestyle and living condition information such as access to piped water and sewage systems, smoking, drinking and exposure to pesticides. 100% of images have a skin phototype using FST while only 65% of images contain patient sex. The median patient age is 62 years old with a minimum age of 6. During manual review of UFES data, we found 43 images with inconsistent or mixed labeling

(e.g. duplicate images assigned to different ground-truth diagnoses: 6 instances where the ground-truth label switches between *scc* and *bcc* and 7 other instances where the label change impacts designation of the lesion as benign/malignant. All mislabelled examples were removed prior to analysis. Images from the UFES dataset have varying sizes with a maximum of 3,474 x 3,476 pixels and a minimum of 147 x 147 pixels. All files were converted to 8-bit RGB color channels, resized using bilinear interpolation and center-cropped to 299 x 299 pixels before use.

The **Diverse Dermatology Images (DDI)**[26] dataset consists of 656 images from Stanford dermatology clinics in the US captured between 2010 and 2020. All photos were taken on a clinic-issued smartphone and extracted retrospectively from electronic health records. We downloaded images and the associated metadata file from the Stanford AIMI Datasets Azure link (https://stanfordaimi.azurewebsites.net/datasets/35866158-8196-48d8-87bf-50dca81df965). The DDI dataset features over 40 unique diagnoses, including examples from all HAM10000 classes, but 36 others with low incidence in the general population and not contained in the other imaging datasets included in our analysis. (Supplementary Table 2) Metadata for DDI images is sparse with only a unique image identifier, diagnosis, malignancy and skin type information included for each lesion. Notably, sex and unique patient or lesion identifiers were absent. Images had varying sizes with a maximum of 1,914 x 1,424 pixels and a minimum of 163 x 79 pixels. We preprocessed DDI by converting images to 8-bit RGB color channels, resizing using bilinear interpolation and center-cropping to 299 x 299 pixels before use. Authors of DDI performed additional data augmentation including random rotation and vertical flipping of images before training their malignancy prediction model, however we were unable to replicate this process due to lack of access to the model's training code.

SAGE Model Training

We first normalized all input pixel values using mean and standard deviations from the ImageNet dataset. All image preprocessing was completed using pillow (v.10.3.0) and the torchvision (v0.17.2) transforms library. The SAGE model architecture was written in python (v3.10.14) using pytorch[40] (v2.2.2) and consists of an encoder as well as decoder and classifier modules that take the encoder's compressed embedding as input. The decoder contains 6 fully-connected layers, 6 deconvolutional layers and a learned upsampling layer while the classifier contains 5 fully connected layers. We used a ResNet50 SAGE model encoder initialized using weights from ImageNet[41] pretraining and made available through the torchvision model library. An embedding vector of size 32 was used for training the decoder and classifier modules. The model was trained using the train split (90%) from the HAM1000 dataset in two stages with a balanced batch sampler and a batch size of 63. An initial warmup stage trained the encoder and classifier using the combined center loss of the latent space embedding and cross-entropy loss of classification until center loss no longer improved (maximum of 150 epochs). The second stage trained the encoder, decoder and classifier for 150 epochs and used cross-entropy loss as well as the decoder's mean-squared error (MSE) of reconstruction. We utilized a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5} with the AdamW[42] optimizer and trained models using two Nvidia A40 GPUs.

t-SNE Visualization and Plotting

We generated the latent space embeddings of size 32 for images in all datasets using the trained SAGE model. For visualization purposes only, these embeddings were mapped to t-stochastic neighbor embedding (t-SNE) space using scikit-learn's manifold library, labeled according to the dataset of origin and plotted as scatterplots using matplotlib[43] (v3.9.0). To explore cluster identities, we plotted HAM10000 train and test images together and colored scatter points according to their diagnostic category. All plots were created using a combination of matplotlib and seaborn[44] (v0.13.2), and figures were assembled using BioRender.

SAGE Scoring

Before evaluation, we preprocessed all images with the same steps taken for the training data and used the trained SAGE model to generate three output measures as previously described and detailed in Fig. 2.[32]

Let $x = (x_1, x_2, x_3)$ represent the observed output values for a given image, corresponding to the three SAGE model measures:

1. x_1: L1 latent embedding distance to the $k = 25$ nearest training neighbors
2. x_2: Softmax classifier confidence (argmax)
3. x_3: Reconstruction error

Let X_i be the random variable denoting the distribution of the model's output for measure i across the training data and let $F_i(x) = P(X_i \leq x_i)$ denote the cumulative distribution function (CDF) of X_i.

We define the **exceedance probability** for measure i as:

$$E_i(x_i) = P(X_i > x_i) = 1 - F_i(x_i)$$

The **SAGE score** for an image is then computed as the **geometric mean** of the three exceedance probabilities:

$$SAGE(x) = \left(\prod_{i=1}^{3} E_i(x_i)\right)^{1/3}$$

Image Annotation and Parallel Coordinates Plot

We manually-annotated 12 independent image features (e.g. lesion contrast to skin, presence of non-skin background) that could impact quality and reliability of malignancy prediction as detailed in Supplementary Table 3 for HIBA, UFES and DDI datasets. Images were reviewed by a single consistent observer (W.M.S.) with clinician oversight (E.B. and R.F.T.). Biasing information such as diagnosis, FST and malignancy status were removed prior to annotation. All annotations were appended to the metadata accompanying each dataset. After annotation was

completed, we normalized the observed levels of 5 image features (lesion contrast, camera flash, hair level, presence of ruler and non-skin background) and plotted a line for each test image with the color denoted by the image's calculated SAGE score. We plotted images in order from low to high-scoring examples and added random jitter of 1.25% to improve visibility of line overlays.

Comparing SAGE Score Across FST Levels for High-Quality Images

Using the manually annotated image features, we removed images from the entirety of the HIBA, DDI and UFES datasets (*n*=3,889) with low lesion pigmentation (contrast ≤ 1), camera flash, measuring devices, high hair density (hair level ≥ 2) and any non-skin background (background ≥ 1) leaving 2,168 high-quality images. We then partitioned these images into groups according to skin-tone levels (FST I-II, III-IV, V-VI) and compared the pairwise distributions of SAGE score values using a two-sided Wilcoxon rank-sum test and a significance threshold of 0.05.

Malignancy Prediction

We used a binary (i.e. malignant v. benign) deep-learning predictor from Daneshjou et al.[26] pre-trained on HAM10000 images and implemented in python according to instructions on the project's GitHub repository, as this was the only model from the paper with a publicly-available training dataset (https://github.com/DDI-Dataset/DDI-Code). Preprocessing steps used before malignancy prediction were the same as for training and evaluating our SAGE model. Only basal cell carcinoma and melanoma classes were considered malignant for HAM10000 images whereas HIBA and UFES also contained examples of malignant squamous cell carcinoma lesions. The DDI dataset came pre-annotated with histologically confirmed malignancies which were used as ground truth labels. All images with an output greater than or equal to the predetermined threshold of 0.733 were classified as malignant, with those falling below classified as benign lesions for all datasets. We calculated intrinsic malignancy prediction confidence as a function of the distance to the decision boundary using the following equation:

$$Confidence = \frac{1}{1 + e^{1(x-0.733)}}$$

Overall Prediction Performance Evaluation

After SAGE scoring and malignancy prediction, a random sample of images (*n* = 500) was taken from the HAM test, HIBA, UFES and DDI datasets and merged to form a mixed test set. Each random sample contained malignant and benign images in a 1:3 ratio mirroring that of the training data and what is encountered in clinical practice.[3] We used a series of progressive SAGE score values ([0.0, 0.1, 0.2, 0.3, 0.4]) and removed images falling below these thresholds from the HAM train and mixed test sets, plotting the proportions of each dataset remaining. At each threshold we calculated the area under the receiver operating characteristic (AUROC) curve using scikit-learn[45] (v1.4.2) to visualize changes to overall performance. To measure changes to positive predictivity, we used the PrecisionRecallDisplay command to plot precision recall (PR) curves and calculate average precision (AP) values after SAGE score thresholding.

We repeated this process using prediction confidence instead of SAGE score and a series of confidence thresholds of [0.0, 0.5, 0.65, 0.8, 0.95]. For score thresholding on image type and skin type, this process was repeated using SAGE thresholds of [0.0, 0.01, 0.05, 0.1, 0.2] and all (unsampled) test images after removal of rows with missing metadata values.

Prediction Performance of Ground-Truth Malignancies

We grouped images in the sampled test dataset with a ground-truth malignant diagnosis by disease. For each diagnostic group, we repeated the process of SAGE score thresholding using a progressive set of values [0.0, 0.01, 0.05, 0.1, 0.2] and removed images falling below each level, plotting the proportion of images remaining. We plotted changes to malignancy prediction accuracy over the same thresholds, with each diagnostic group represented by a distinct line. We then plotted a joint grid of the raw malignancy prediction values against the calculated SAGE scores, isolating the false negative (FN) predictions falling below the malignancy predictor's threshold of 0.733. We used a split bar plot to show enrichment of FN examples for each test diagnostic group below a conservative SAGE score threshold (0.151) where 90% of the HAM10000 training dataset is retained.

Code & Data Availability

All datasets used in our analysis are publicly available and download links are included in Methods. Code for SAGE model training and retrieving SAGE scores is freely available on our GitHub repository (https://github.com/pdxgx/sage).

Acknowledgements

This study was supported by a VA Career Development Award to R.F.T. (1IK2CX002049-01). The contents do not represent the views of the US Department of Veterans Affairs or the US Government. No terms were censored from this text.

Figures

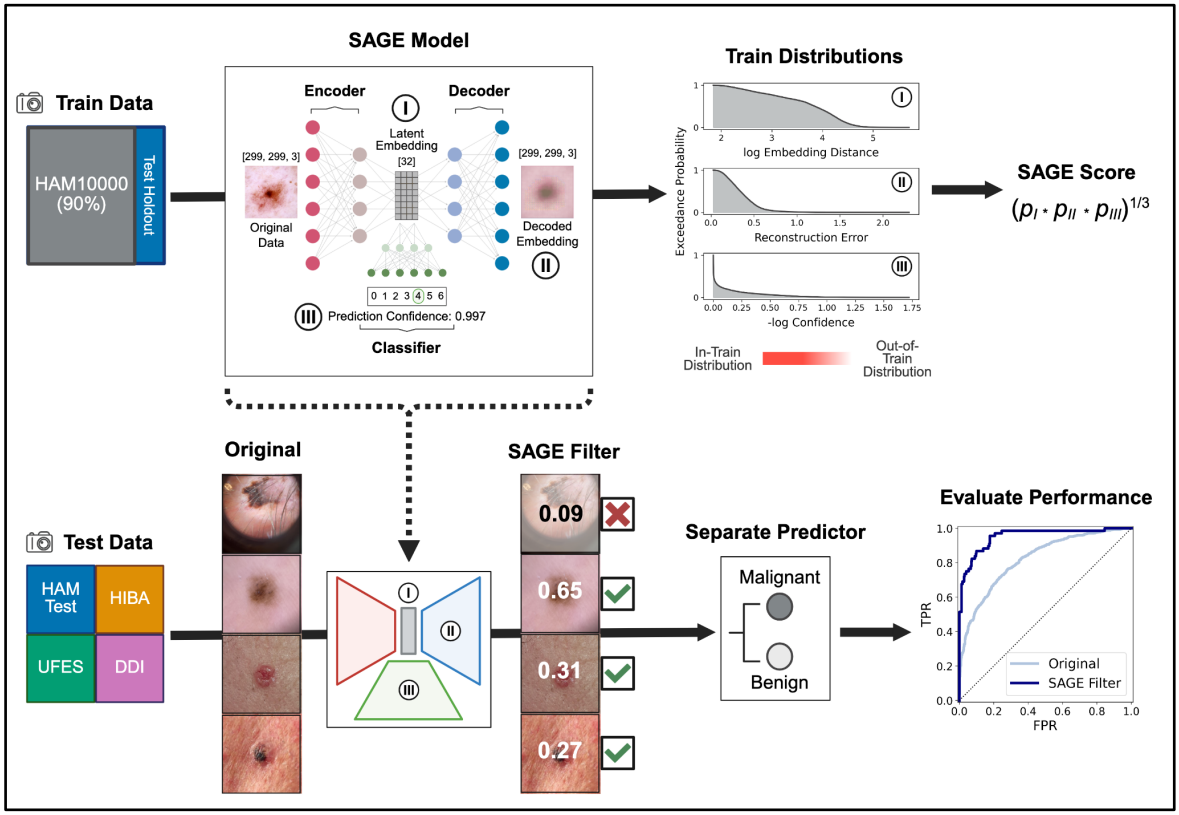


Figure 1 – Overview of SAGE model training and deployment for improving performance of a separate malignancy predictor. In the top-left corner, the HAM10000 dataset is split into training and test sets with the training set used to fit the SAGE model's encoder, decoder and classifier components. The location of the three SAGE components are shown within the model's greater architecture with each component assigned a label (I-III). To the right, the output values for each component is shown as a labeled distribution corresponding to exceedance probabilities of the training data (Methods). Values move from more likely ("In-Train Distribution") to less likely ("Out-of-Train Distribution") to be encountered in the training data as points move from left to right along the x-axis. The SAGE score is shown as the geometric mean of the three exceedance probabilities. In the bottom-left corner, 4 images from the test datasets are shown before ('Original') and after ('SAGE Filter') scoring with the trained SAGE model. Boxes with checkmarks show images with SAGE scores above a cutoff of 0.2, with a red 'X' denoting where scores fall below this amount. A separate, pre-trained model is used to predict malignancy of test images and performance is evaluated via an example Receiver Operating Characteristic (ROC) curve. The original and SAGE score-filtered images demonstrate improvement after low-scoring images are removed from the pool of predicted examples as assessed by area-under the ROC curve.

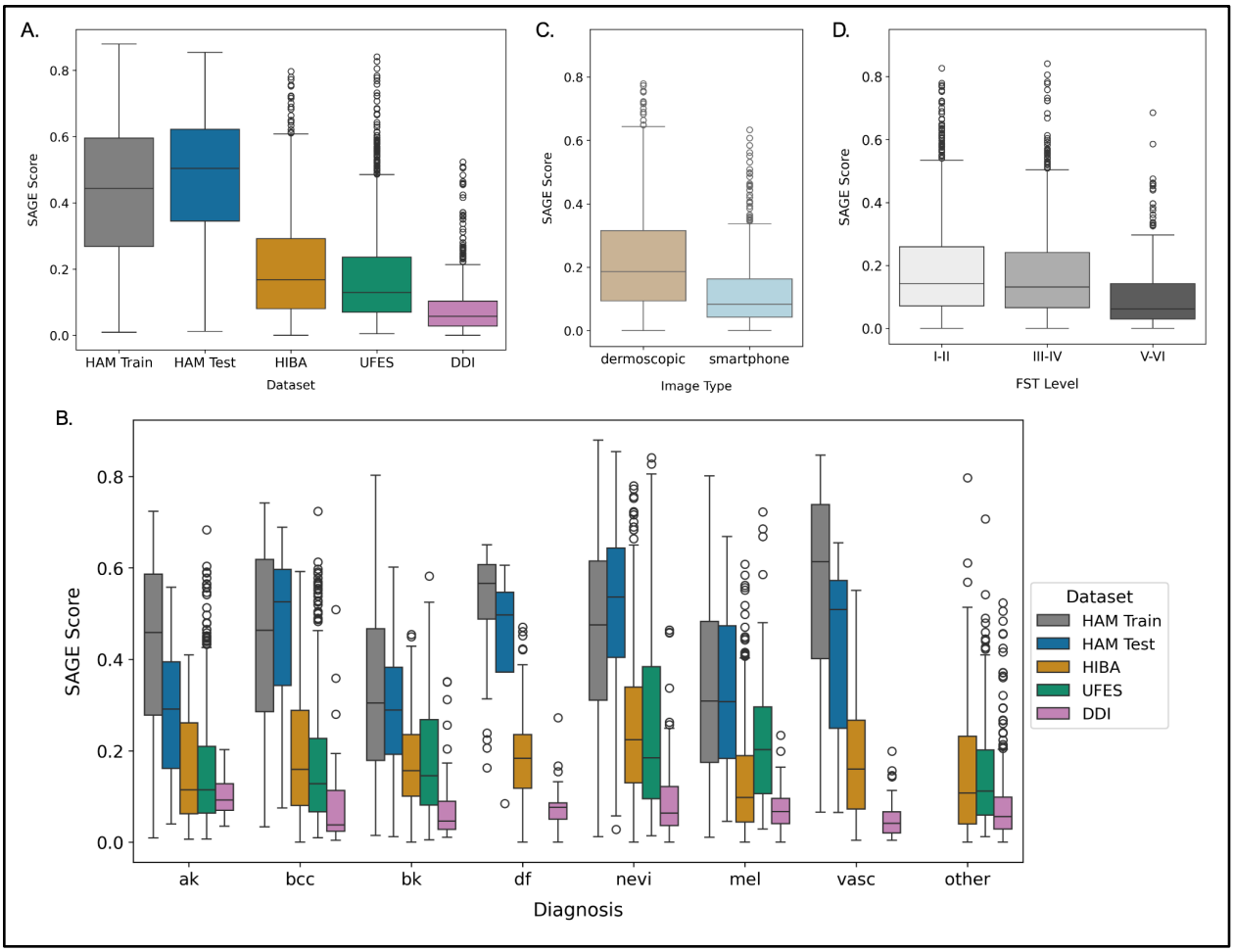


Figure 2 – SAGE scores reveal differences between and within dataset categories. A) Distributions of SAGE scores for imaging datasets calculated using the model fit to the HAM10000 training split are represented as a uniquely-colored boxplot, with the central line denoting the median score values. Top and bottom of each box corresponds to the 25th and 75th percentile of values and whiskers extend to 1.5x the interquartile range (IQR). B) SAGE score shown for each of the seven HAM10000 diagnostic categories and split by dataset. Boxes use the same dataset-specific coloring as in A and categories missing from datasets are omitted (e.g. dermatofibroma not present in UFES data). Images with a diagnostic category not present in HAM10000 data are grouped together as an 'other' category. C) Test images from HAM10000, HIBA and DDI are split according to dermoscopic or clinical smartphone imaging technology, with SAGE score distributions shown as separate boxplots. D) Images from HIBA, UFES and DDI with skin-tone metadata are split according to three FST level groupings (I-II, III-IV and V-VI), with SAGE score distributions for each skin-tone group represented as a boxplot. (ak –actinic keratosis, bcc – basal cell carcinoma, df – dermatofibroma, nevi – melanocytic nevi, mel – melanoma, vasc – vascular skin lesion)

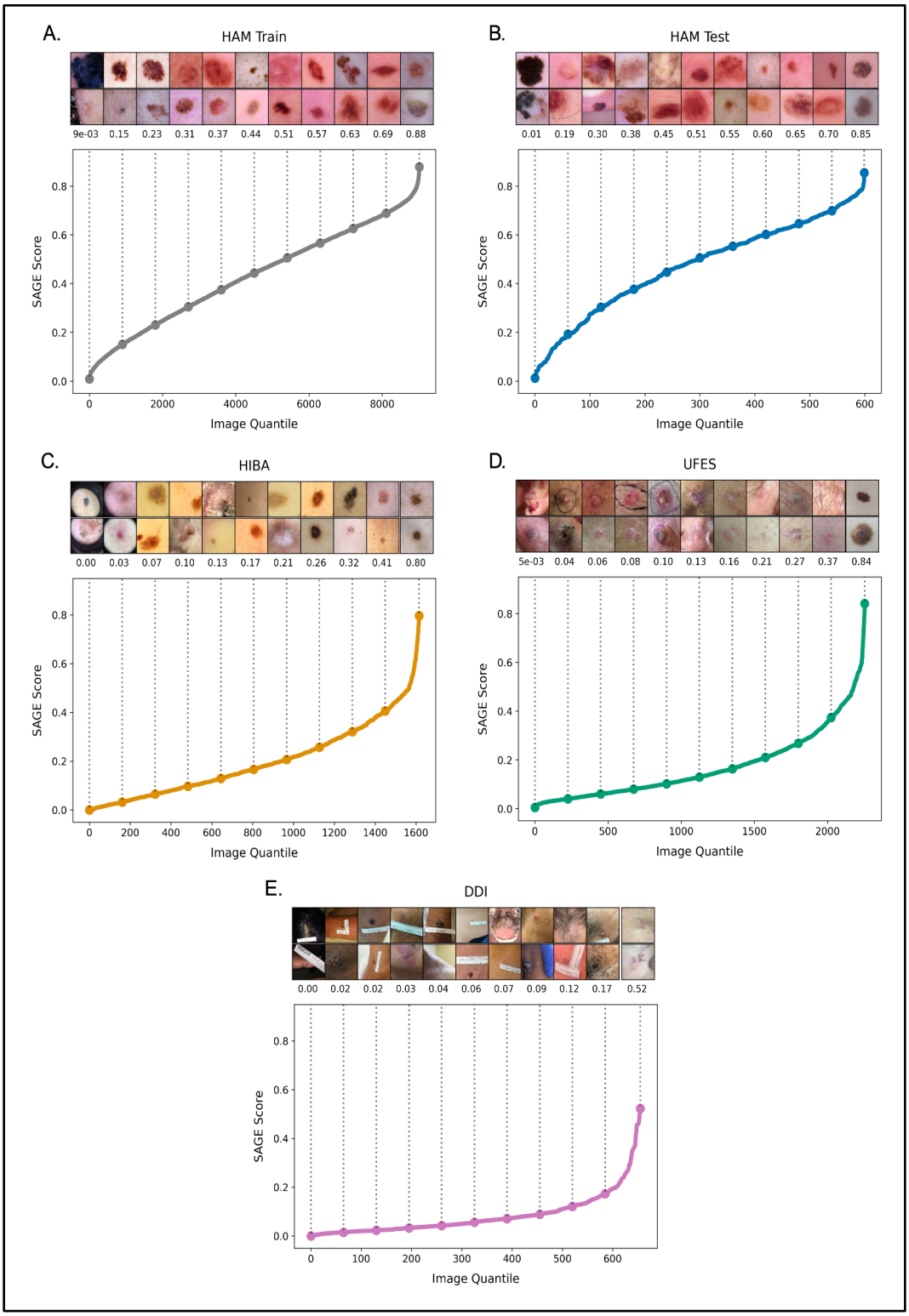


Figure 3 – Quantile plots show imaging artifacts associated with low SAGE scores. Images are sorted by SAGE score (y-axis) and plotted by image quantile (x-axis) for the A) HAM10000 Train (*n* = 9,013), B) HAM10000 Test (*n* = 600), C) HIBA (*n* = 1,616), D) UFES (*n* = 2,255) and E) DDI (*n* = 656) datasets. For each quantile plot a scatter point marks decile intervals and a vertical dotted line leads to

the two nearest SAGE score values and the corresponding lesion images after resizing and center crop transforms have been applied.

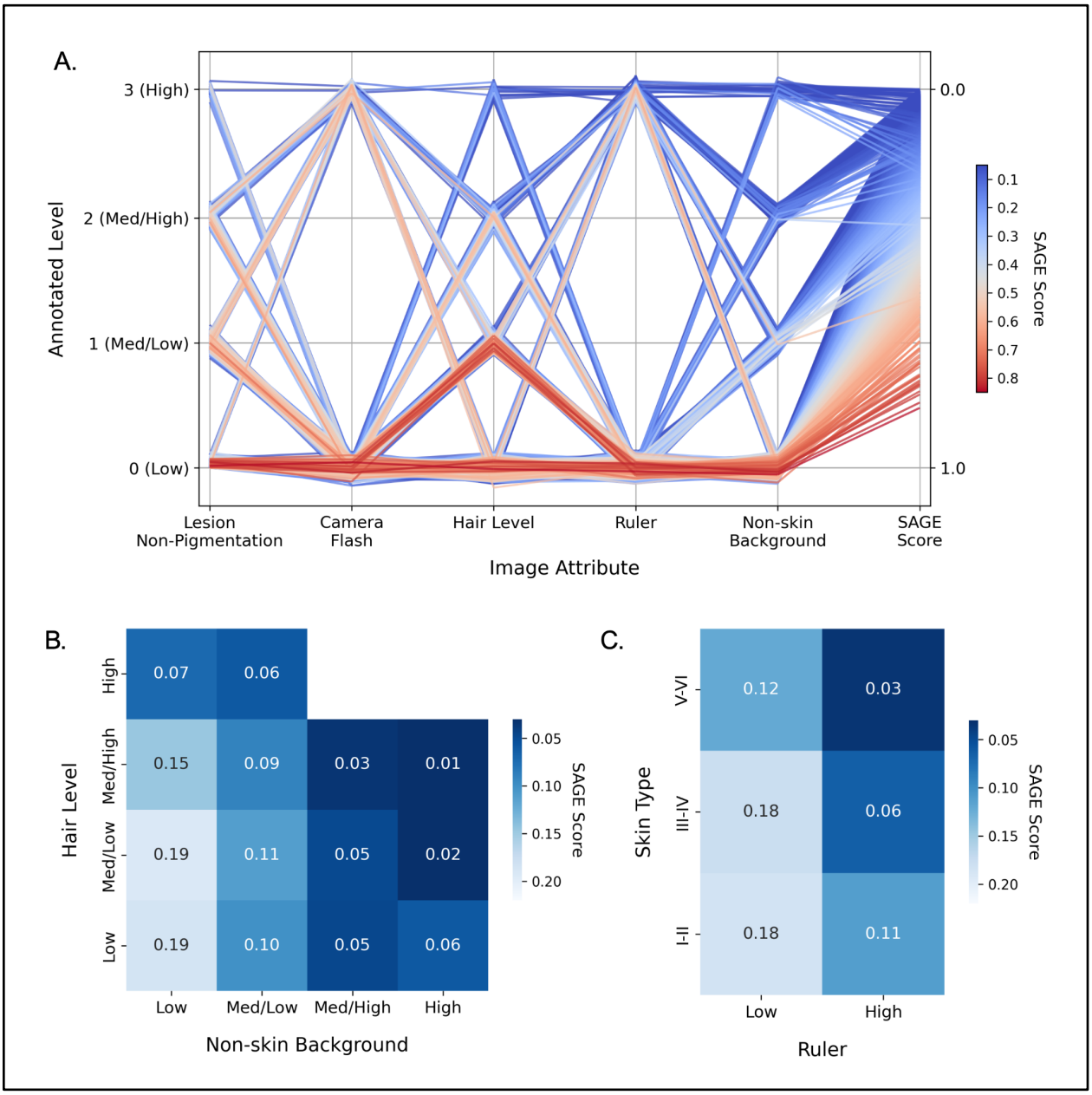


Figure 4 – Increasing severity of manually-annotated image features correlates with decreased SAGE scores. A) Five manually-annotated attributes for HIBA, UFES and DDI images (*n*=3,889) are shown along the x-axis with normalized level plotted on the left y-axis, with the right y-axis showing SAGE score from low (top) to high (bottom). Each image's combination of features is represented as a continuous line and is assigned a color based on its SAGE score value, with intensity determined by the colorbar gradient to the plot's right. B) Heatmap of SAGE score by manually-annotated hair density (y-axis) and non-skin background (x-axis) feature severity. Squares are colored based on the average

SAGE score value shown in the center of each square, with color intensity determined by the colorbar gradient to the right of the heatmap. Each row corresponds to one severity level of hair density and displays the change in average SAGE score as non-skin background increases from left to right. C) Heatmap of SAGE score by skin type and measuring device levels. Test images are grouped into three categories based on FST into three levels (I-II, III-IV and V-VI) and displayed along the y-axis. The x-axis displays the binary presence of a measuring device (ruler) in the lesion image, with heatmap squares displaying the average SAGE score for images at the intersection of the two categories.

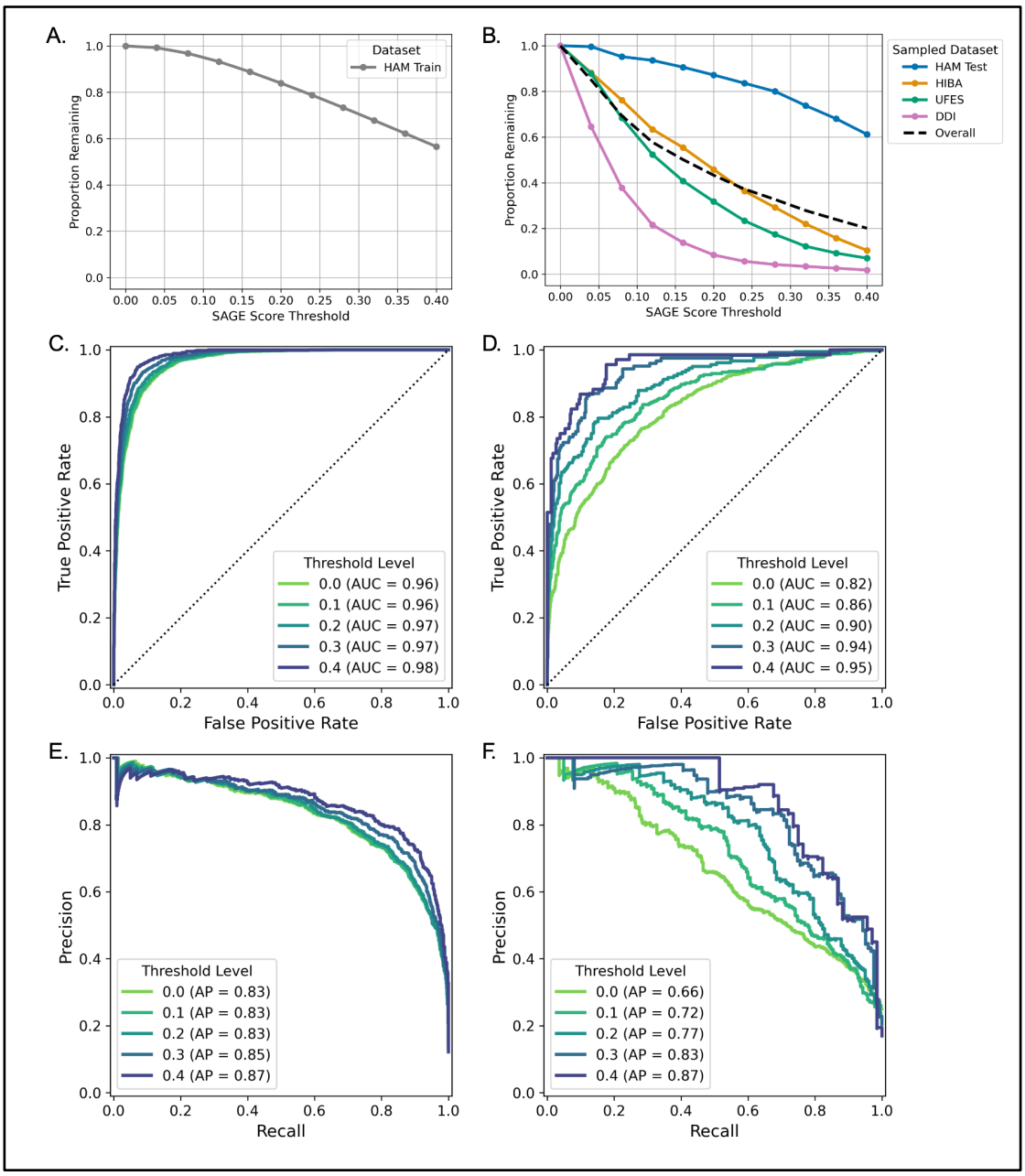
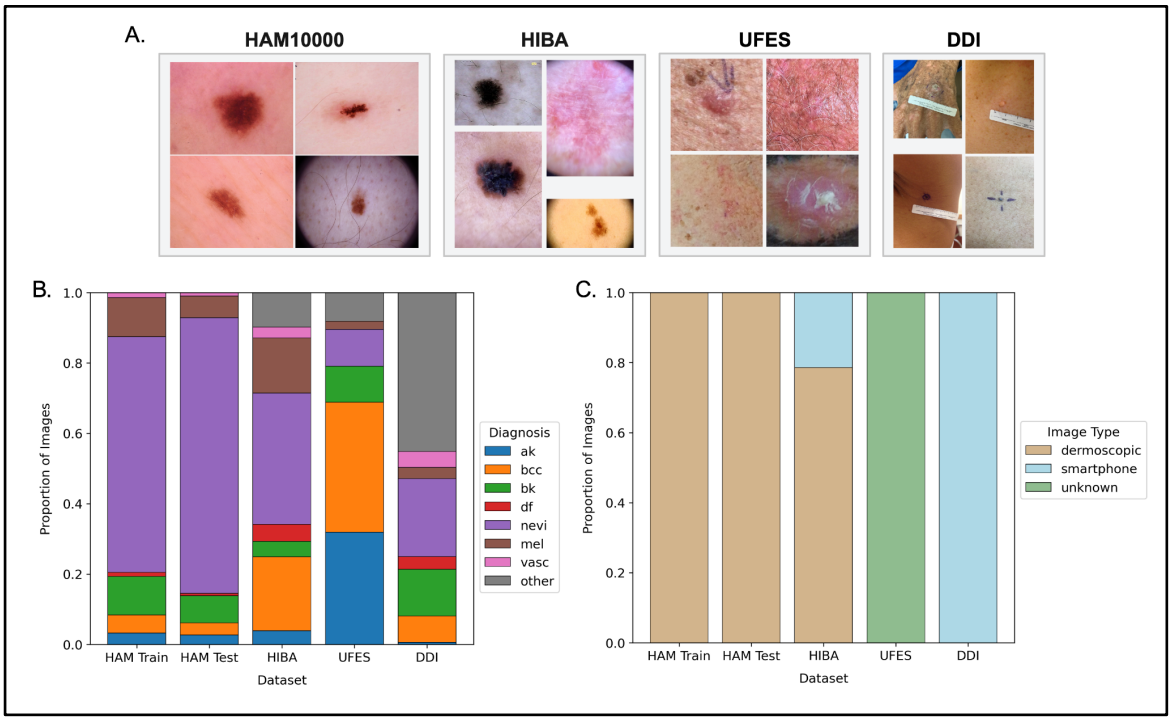


Figure 5 – SAGE score filtering improves performance of a pre-trained malignancy prediction model. A-B) Line plots for HAM10000 train and a balanced sample of test images (*n*=500 per test dataset) show proportional rates of decrease as images below a series of SAGE score thresholds are removed. The x-axis shows the score thresholds from [0.0 - 0.4] with the y-axis denoting the proportion of the original images remaining after pruning at the corresponding score level. The overall decrease of the proportion of test images remaining is represented as a dotted line. C-D) The remaining images at five SAGE score thresholds undergo malignancy prediction with a separate pre-trained model for the HAM10000 training and balanced test set. The raw prediction values are used to generate receiver operating characteristic (ROC) curves for each dataset, with a new curve generated for each SAGE score threshold level. A dotted line is plotted representing performance of a model with no-skill, and the area-under the curve (AUC) is calculated and shown next to each threshold level in the ROC legend. E-F) The process is repeated for train and test datasets, with precision and recall (PR) curves demonstrating improvements to positive predictive performance after each round of SAGE score thresholding. Average precision (AP) is calculated and displayed next to threshold levels in the PR curve legends.

Supplementary Figures

	HAM10000	HIBA	UFES	DDI
Origin	Austria, Australia	Argentina	Brazil	United States
Total Images	10,015	1,616	2,298	656
Unique Patients	*unknown*	623	1,373	570
Unique Lesions	7,470	1,246	1,641	*unknown*
Imaging Technology	Dermoscope	Dermoscope, Smartphone	Dermoscope, Smartphone	Smartphone
Male / Female %	54.30 / 45.70	46.90 / 53.10	49.60 / 50.40	*unknown*
Light Skin % (FST I-II)	*unknown*	80.00	41.60	36.70
Medium Skin % (FST III-IV)	*unknown*	20.00	55.20	31.70
Dark Skin % (FST V-VI)	*unknown*	0.00	3.20	31.60

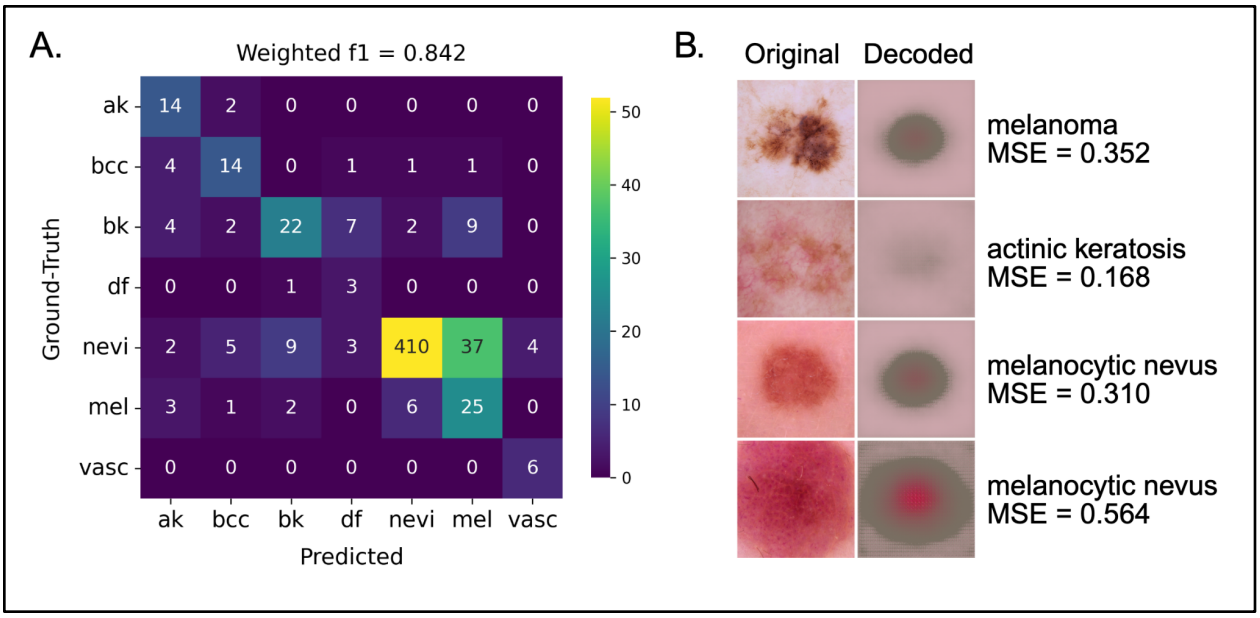
Supplementary Table 1 – Overview of dermatology imaging datasets. Table rows correspond to key metadata categories for each dataset including country of origin, lesion and patient demographic statistics. Each unique dataset is assigned its own column. Missing category information is denoted as "unknown".



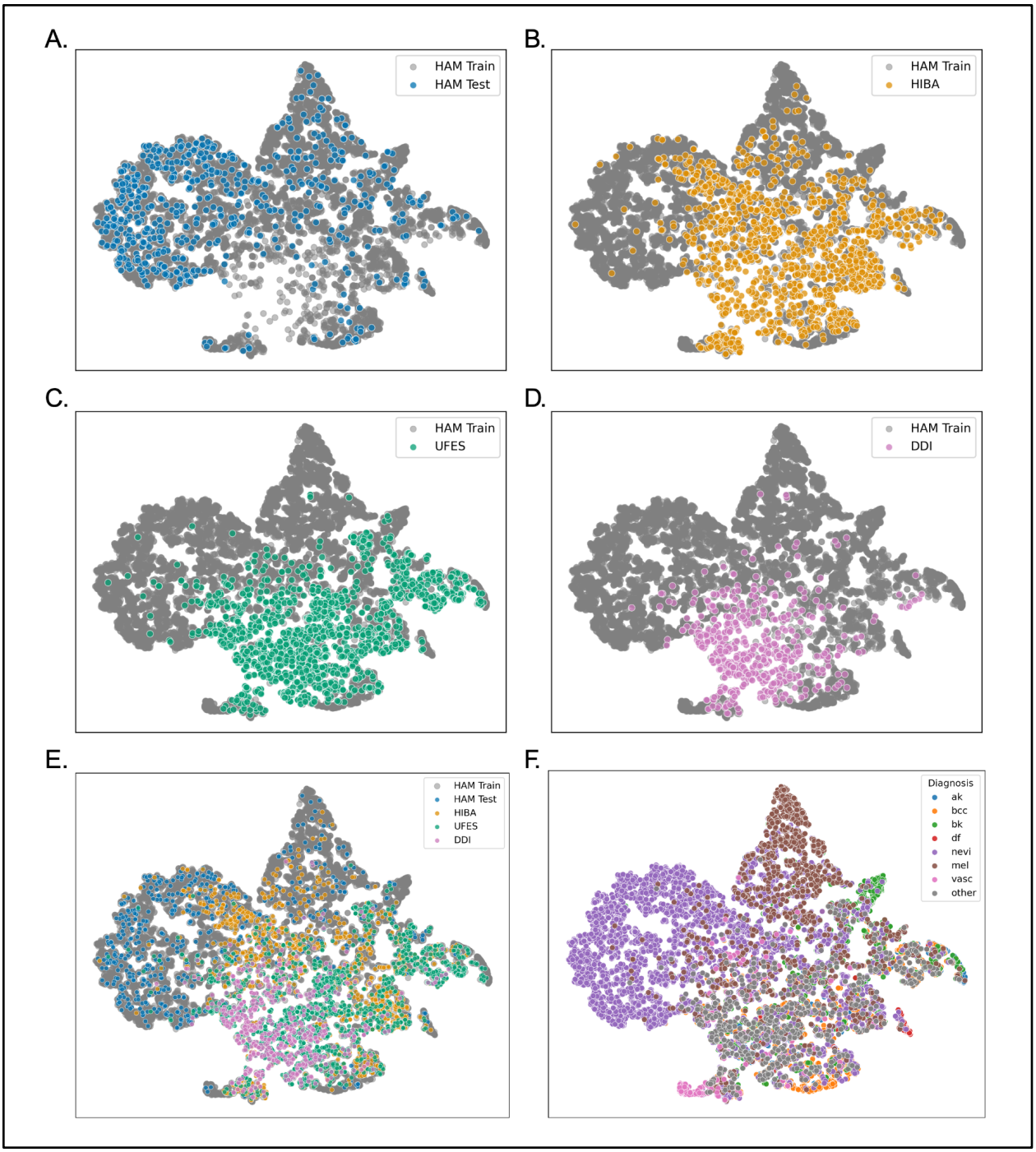
Supplementary Figure 1 – Visualization of differences in dermatology dataset metadata. A) Each unique dataset is represented as a rectangle with a random selection of 4 images showcasing dataset-wide variations in image proportions, technology, image angle. Dermoscopic images from HAM10000 and HIBA show head-on perspectives with lesions centered in the image frame whereas UFES and DDI exhibit a greater variety of lighting and angle changes with the introduction of clinical smartphone imaging. B) Stacked barplots of the proportion of datasets corresponding to diagnostic categories. Each dataset is shown on the x-axis with its own stacked barplot with the proportion of all diagnoses summing to 1. The seven diagnoses from the HAM10000 train and test data are represented by colored bars while additional diagnoses not present in the HAM10000 dataset are shown as a gray 'other' category. C) Stacked barplots showing the proportion of imaging modality present in each dataset. X-axis labels correspond to unique datasets and the total proportion of images sums to 1. For UFES, image type is listed as 'unknown' due to authors failing to designate the technology used for each corresponding image. (ak –actinic keratosis, bcc – basal cell carcinoma, df – dermatofibroma, nevi – melanocytic nevi, mel – melanoma, vasc – vascular skin lesion)

Label	Diagnosis	Label	Diagnosis
7	Squamous cell carcinoma	26	Eccrine-poroma
8	T-cell lymphomas and leukemias	27	Keloid body
9	Lipoma	28	Hematoma
10	Hyperpigmentation	29	Metastatic carcinoma
11	Acochordon	30	Folliculitis
12	Verruca vulgaris, Verruciform xanthoma	31	Acne
13	Foreign body granuloma	32	Xanthogranuloma
14	Abbrasion	33	Eczema
15	Epidermoid cyst	34	Molluscum-contagiosum
16	Acquired digital fibrokeratoma	35	Condyloma-accuminatum
17	Trichilemmoma, Trichofolliculoma	36	Morphea (localized scleroderma)
18	Neurofibroma, Neuroma, Cellular-neurothekeoma (CNT)	37	Dermatomyositis
19	Syringocystadenoma papilliferum (SCAP)	38	Graft-versus-host disease
20	Nevus lipomatosus superficialis (NLS)	39	Coccidioidomycosis (Valley Fever)
21	Onychomycosis	40	Chondroid-syringoma
22	Lymphocytic hypophysitis	41	Tinea pedis
23	Prurigo-nodularis	42	Clear cell acanthoma
24	Kaposi sarcoma	43	Abscess
25	Scarring		

Supplementary Table 2 – Lesion diagnoses missing from HAM10000 dataset. Table of diagnoses present in the HIBA, UFES and DDI datasets that are not present in the HAM10000 training and test data. Labels start from the end point of the seven HAM10000 classes [0-6]. Class 7, squamous cell carcinoma, is present in all three test datasets mentioned whereas classes 8-43 are only present in DDI. Where applicable, similar diagnoses are grouped into the same label category such as wart-like lesions (12) and benign nerve cell tumors (18).



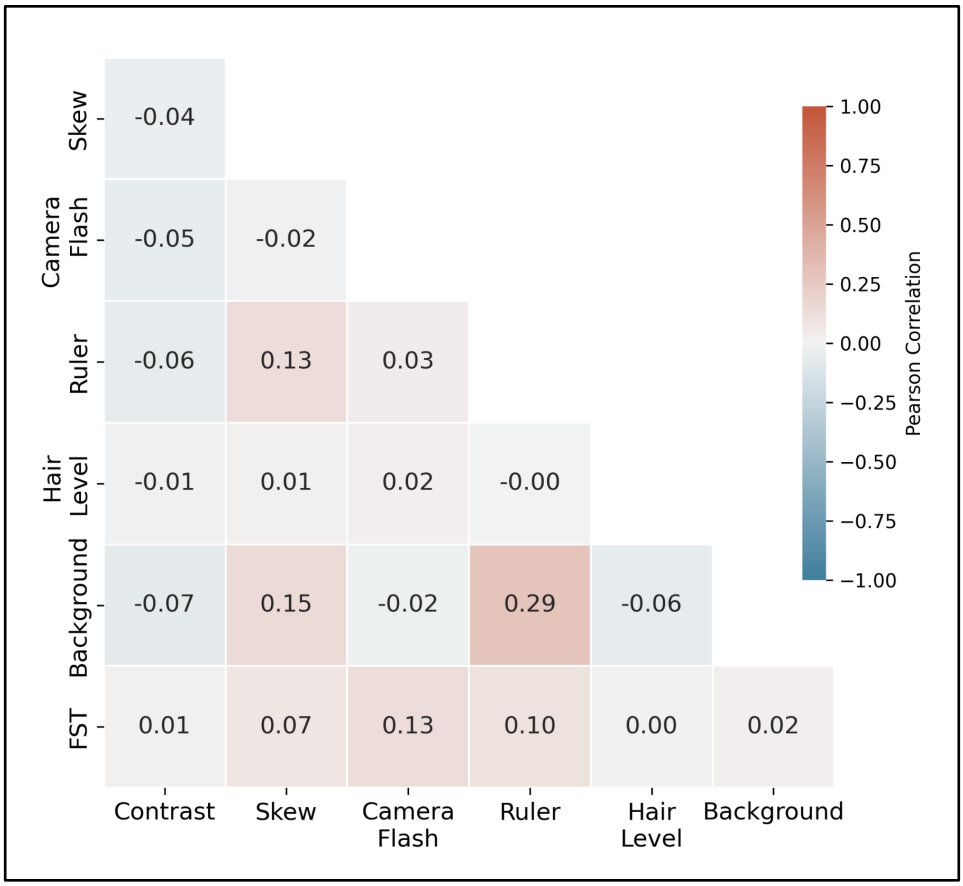
Supplementary Figure 2 – Performance of SAGE model classifier and decoder on HAM10000 test images. A) Confusion matrix of trained SAGE model classifier predictions (x-axis) versus ground truth diagnostic labels (y-axis). Each row displays the predicted classes for a ground-truth label, with squares showing the number of images for each predicted class. Squares are colored according to the values in the colorbar to the right of the confusion matrix and the overall weighted F1 score is displayed above the plot. B) Original images from the HAM10000 test set are displayed in the left-hand column with the result of the embedding and decoding process using the trained SAGE model displayed in the right-hand column. For each row, the diagnostic label is displayed as well as the mean-squared error (MSE) of the decoded image.



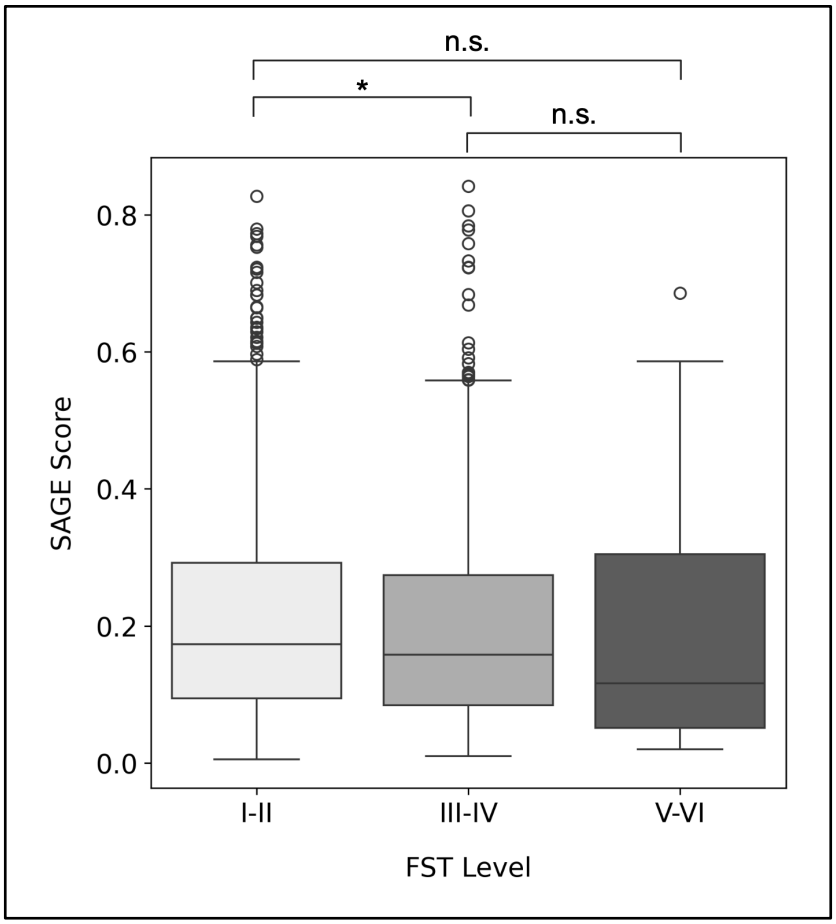
Supplementary Figure 3 – Visualization of 32-dimensional embedding space with t-SNE. A-E) Images for HAM10000, HIBA, UFES and DDI test datasets are compressed using the trained SAGE model encoder, with the resulting embeddings projected into t-SNE space. HAM10000 train image embeddings are plotted in gray to illustrate the reference embedding space after dimensionality reduction with t-SNE. F) Images from all datasets are projected into t-SNE space and colored according to ground-truth label, with clusters arising based on diagnosis such as melanocytic nevi (*nevi*) and melanoma (*mel*). Images with a diagnosis not present in the HAM10000 dataset are labeled in gray as an 'other' category.

Attribute	Description	Variable	Level 0	Level 1	Level 2	Level 3
Size	Lesion size w.r.t. cropped image frame	Ordinal	No discernable lesion	Very small to small size, nearly all skin	Majority of image is skin, lesion visible	Majority of image is lesion, skin visible in background
Contrast	Lesion contrast with skin	Ordinal	No contrast	Lesion barely visible, blends with skin	Lesion clearly visible	Lesion starkly visible against skin background of different color
Hair Level	Hair density present in image	Ordinal	No hair	Some hair visible	Denser hair (arm, chest, partial scalp)	Very dense hair (scalp)
Skew	Off-center position of lesion in image	Ordinal	Lesion centered	Some off-center skew	Lesion on periphery of image or cropped	–
Non-skin Background	Background visible in image	Ordinal	No non-skin background	Any non-skin visible in image background	¼ - ½ of image consists of non-skin background	Over ½ of image is non-skin background (dermoscope border, clinic)
Cover	Cloth or bandage covering of skin	Ordinal	No skin covering	Any skin covering visible	¼ - ½ of skin is covered	Over ½ of skin in image is covered (hospital gown, bandage)
Marking	Colored marker on skin denoting lesion location	Binary	No marking	Marking present	–	–
Ruler	Measuring device in image	Binary	No ruler	Ruler present (white, black, green)	–	–
Flash	Camera flash used	Binary	No flash	Flash used	–	–
ENEM	Eyes, nose, ear or mouth visible in image	Binary	Not present	Present	–	–
Blurry	Unfocused image	Binary	Image focused	Image unfocused	–	–
Multiple	More than one lesion of similar size present	Binary	One lesion	Multiple lesions	–	–

Supplementary Table 3 – Description of manually-annotated image features. Rows correspond to 12 independent features used to manually-annotate all HIBA, UFES and DDI images. Each row contains a brief written overview, the variable type and descriptions of up to 4 levels of severity for ordinal variables and 2 levels for binary variables.

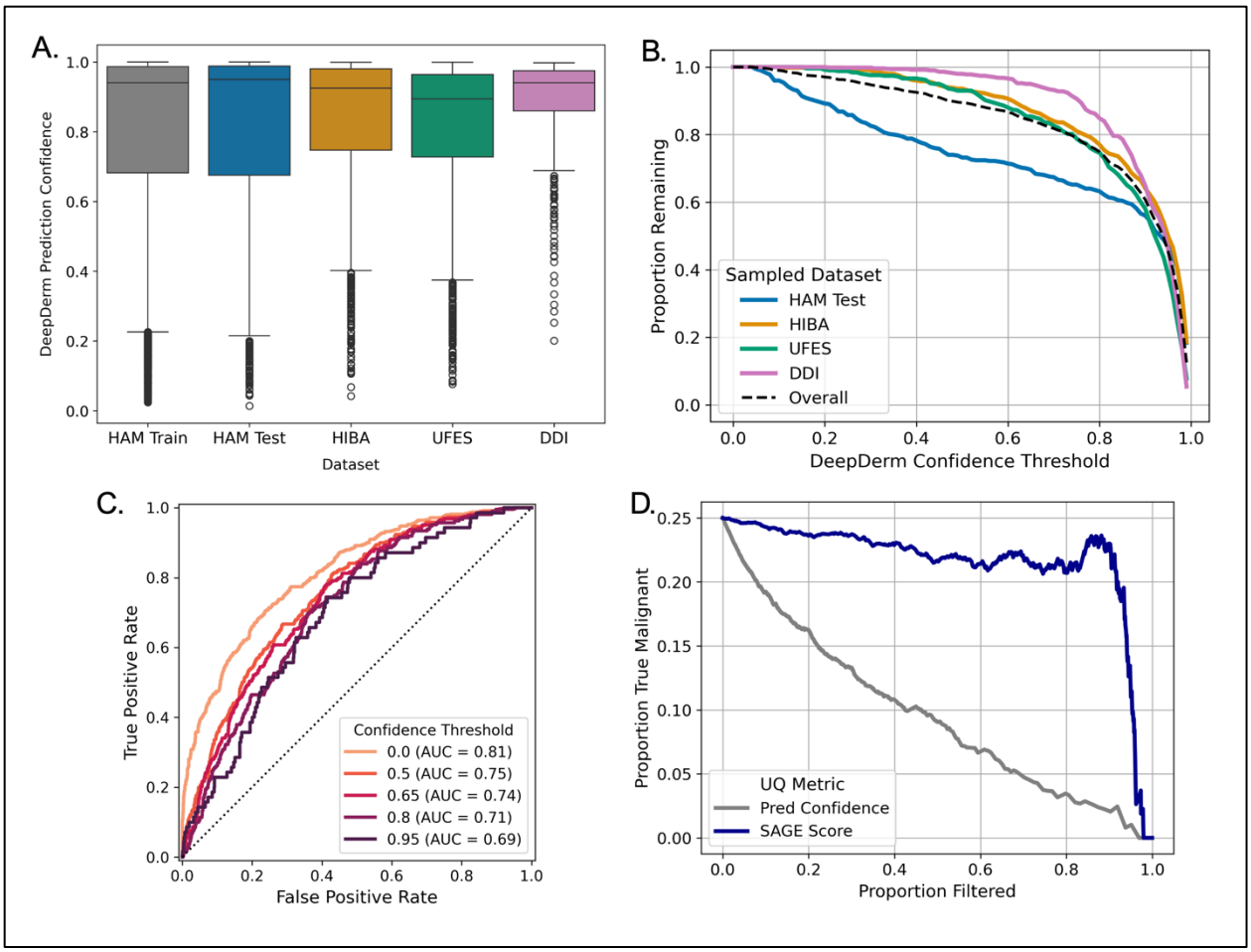


Supplementary Figure 4 – Correlation analysis of manually-annotated image features. A lower triangular matrix displays the Pearson correlation values of 7 unique manually annotated image features for the combined HIBA, UFES and DDI datasets. Each square contains the calculated correlation coefficient and a color based on the gradient shown in the color bar to the right of the matrix. Positive coefficients signify a positive relationship whereas a negative sign shows a negative relationship, with values under 0.3 considered to be a weak correlation.

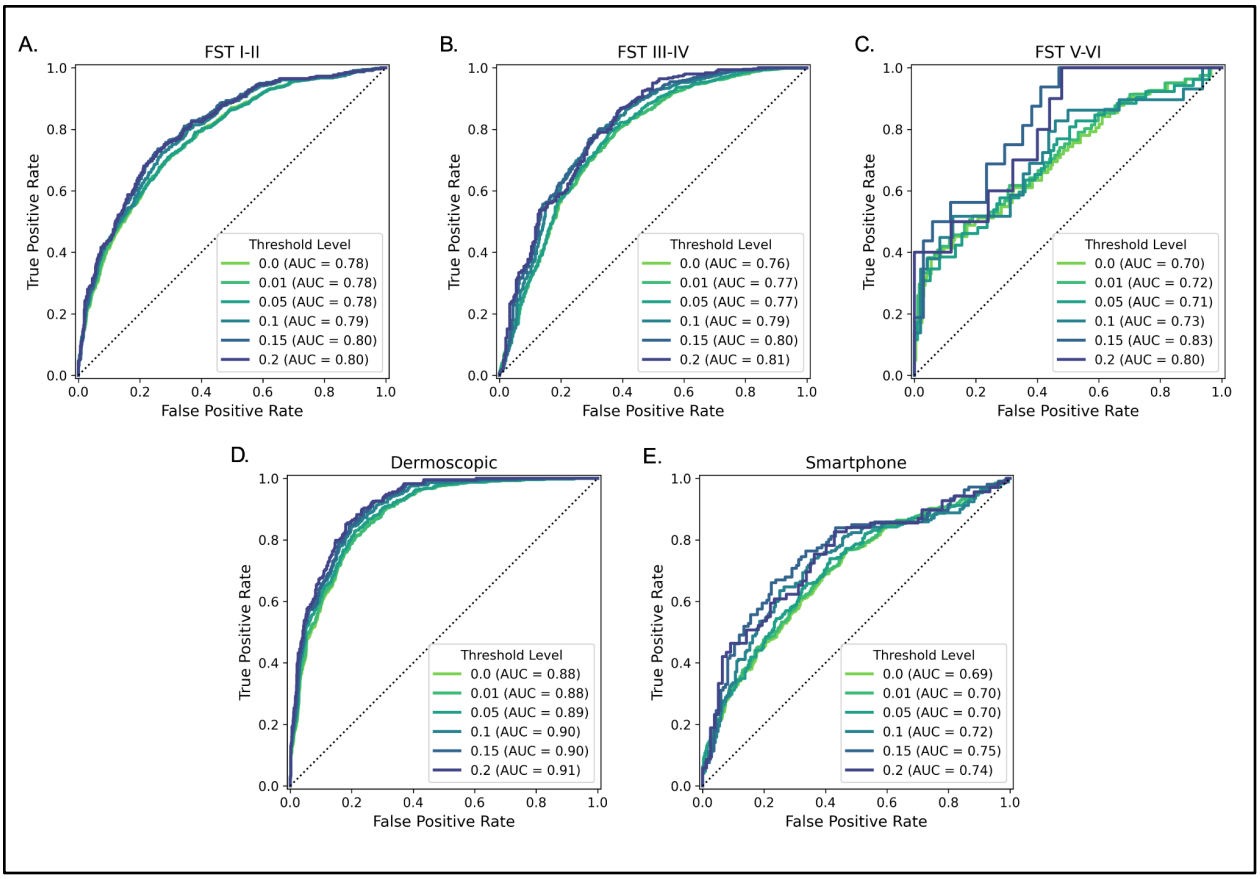


Supplementary Figure 5 – SAGE score distributions for FST levels display overlap after filtering low-quality images. SAGE score distributions for HIBA, UFES and DDI datasets are combined and split by FST level after filtering examples with manually-annotated features indicative of low image quality (Methods), with distributions shown as boxplots and statistical significance between distributions calculated using the Wilcoxon rank-sum test ($\alpha = 0.05$). SAGE score differences are not significant (n.s.) between FST levels I-II and V-VI (p-value = 0.06) and levels III-IV and V-VI (p-value = 0.19) and significant between levels I-II and III-IV (p-value = 0.03).

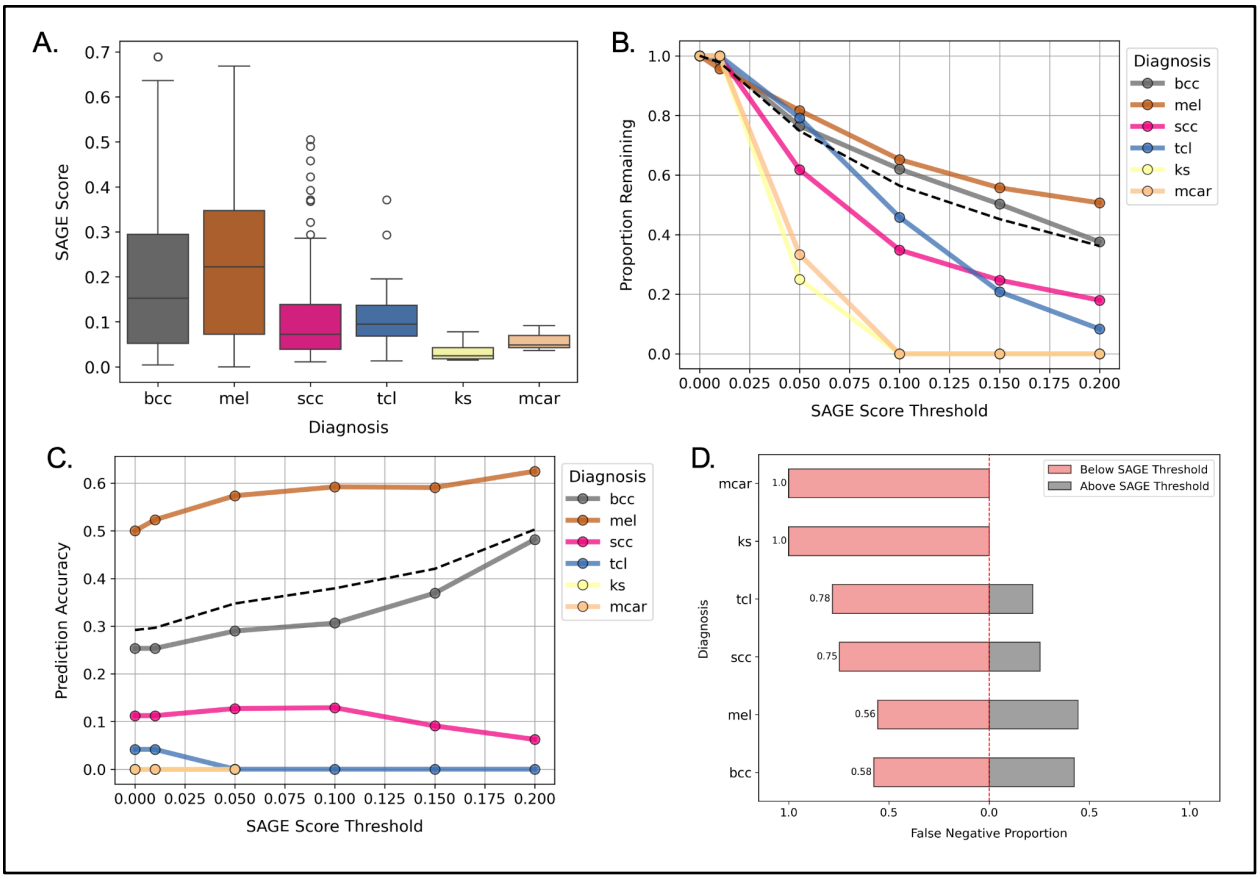
S Fig. 6



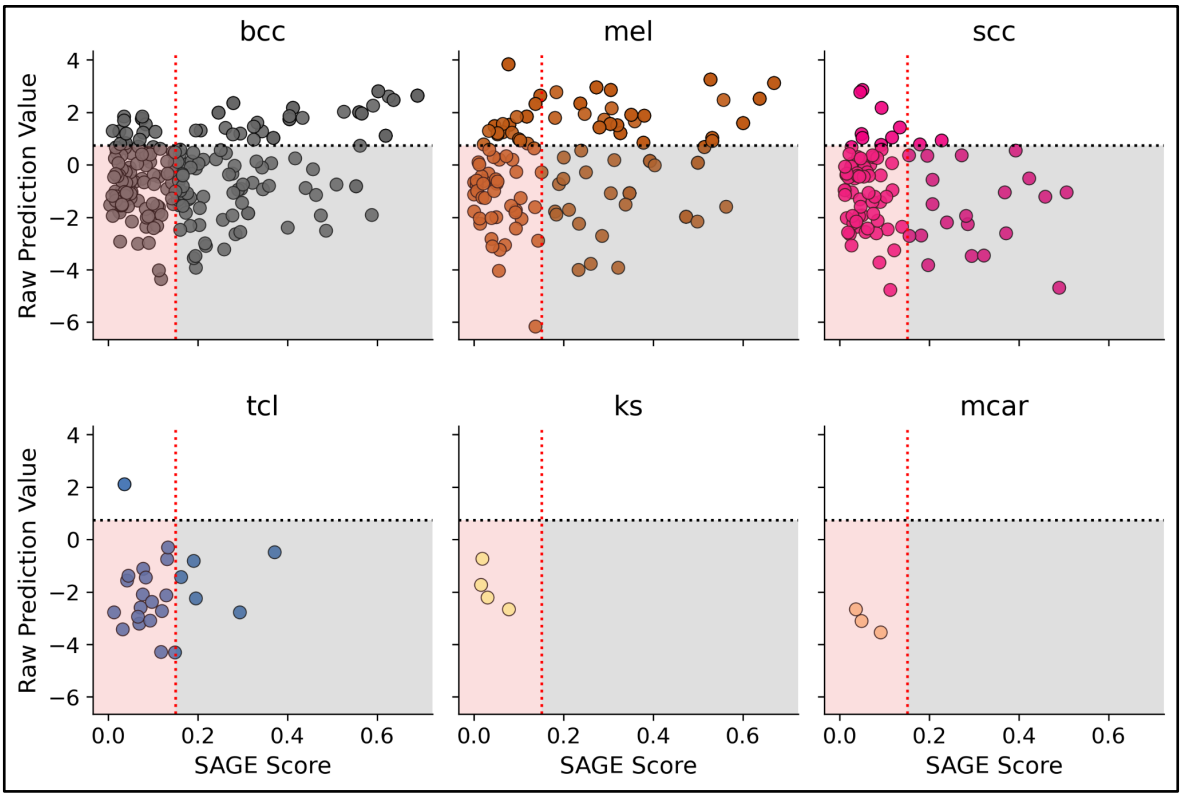
Supplementary Figure 6 – Model-intrinsic confidence thresholding does not improve malignancy prediction. A) Box plot of distribution of malignancy confidence scores derived from the pretrained model predictions on the y-axis with datasets shown on the x-axis. B) Proportion of images remaining (y-axis) when samples below model intrinsic confidence score thresholds (x-axis) are removed. Each line plot represents a different random sample of images from a test dataset (*n*=500) with a consistent 1:3 ratio of malignant to benign labels. C) Receiver operating characteristic (ROC) curves for the sampled test dataset after filtering images below five increasing thresholds of malignancy prediction confidence. Area-under-the-curve (AUC) is displayed alongside each confidence threshold value in the figure legend. D) The proportion of true malignant samples (y-axis) as increasing thresholds of intrinsic malignancy prediction confidence (gray) or SAGE score (navy) are used to filter out-of-distribution images. The x-axis indicates the proportion of images that are removed at increments of 0.001 from 0.0 to the maximum threshold value.



Supplementary Figure 7 – ROC curves show improvement of separate malignancy predictor performance for key metadata categories. A-C) Test images for HIBA, UFES and DDI are split by FST level and ROC curves are shown after filtering examples below five SAGE score threshold values [0.0 - 0.2]. A distinct ROC curve is plotted for each SAGE score threshold across the separate malignancy predictor's raw output values, with the false positive rate shown on the x-axis and true positive rate displayed on the y-axis. Area-under the ROC curve (AUC) is displayed beside the value in the legend. Curves become darker as SAGE score threshold values increase. D-E) Test images for HAM10000, HIBA and DDI are split by imaging technology and filtered below the same SAGE score levels, with ROC curves generated using retained images above each threshold.



Supplementary Figure 8 – Malignant disease classes missing from HAM10000 are enriched for low SAGE scores. A) Images from the sampled test dataset with ground-truth malignant diagnoses are grouped by disease label and SAGE score distributions are plotted as boxplots for each group. Distributions of basal cell carcinoma and melanoma lesions contain examples from the HAM10000 test dataset whereas the other diagnoses are not featured in the training or test data. B) Separate line plots for each malignant disease class show the proportional loss of images on the y-axis as they are filtered at a series of SAGE score thresholds [0.0 - 0.2] displayed on the x-axis. A dashed line shows the proportion of all malignant test images over the SAGE score thresholds. C) Line plots show the changes to accuracy of a separate, pre-trained malignancy predictor after images at low SAGE score thresholds are removed, with each disease class from the sampled test set shown as a separately colored line. A dashed line shows the overall change in prediction accuracy across the increasing SAGE score thresholds. D) False negative (FN) predictions for each disease class from the sampled test dataset below the cutoff value of 0.733 are split into two groups: those above and below a SAGE score threshold of 0.151 where 90% of the HAM10000 training data is retained. The x-axis is centered at 0.0 and the red bars left of the center line show the absolute proportion of FN examples falling below the SAGE score threshold for each disease class. Bars to the right of the center line show the proportion of FN images that are above the SAGE score threshold in grey. (bcc – basal cell carcinoma, mel – melanoma, scc – squamous cell carcinoma, tcl – T-cell lymphomas, ks – kaposi sarcoma, mcar – metastatic carcinoma)



Supplementary Figure 9 – Scatterplot display of malignancy prediction and SAGE scoring for cancerous test lesions. Ground-truth malignant diagnoses from the sampled test dataset are shown as a unique scatterplot in a joint grid, with points colored according to diagnosis. The y-axis values correspond to the raw outputs of downstream malignancy prediction and the x-axis displays the SAGE score for each image. The horizontal dotted black line denotes the cutoff for positively predicting an image as malignant (0.733) while the vertical red line displays the SAGE score value at which 90% of the HAM10000 training data is retained (0.151). A transparent red rectangle shows where false negative images fall below the prediction threshold and the SAGE score cutoff, where the gray rectangle shows false negative lesions that pass the SAGE score cutoff. (bcc – basal cell carcinoma, mel – melanoma, scc – squamous cell carcinoma, tcl – T-cell lymphomas, ks – kaposi sarcoma, mcar – metastatic carcinoma)

References

1. Langselius, O. *et al.* Global burden of cutaneous melanoma incidence attributable to ultraviolet radiation in 2022. *Int. J. Cancer* **n/a**,.
2. Zhou, L., Zhong, Y., Han, L., Xie, Y. & Wan, M. Global, regional, and national trends in the burden of melanoma and non-melanoma skin cancer: insights from the global burden of disease study 1990–2021. *Sci. Rep.* **15**, 5996 (2025).
3. Viola, K. V. *et al.* Outcomes of Referral to Dermatology for Suspicious Lesions: Implications for Teledermatology. *Arch. Dermatol.* **147**, 556–560 (2011).
4. Goyal, M., Knackstedt, T., Yan, S. & Hassanpour, S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Comput. Biol. Med.* **127**, 104065 (2020).
5. Vaidya, T., Zubritsky, L., Alikhan, A. & Housholder, A. Socioeconomic and geographic barriers to dermatology care in urban and rural US populations. *J. Am. Acad. Dermatol.* **78**, 406–408 (2018).
6. Nazari, S. & Garcia, R. Automatic Skin Cancer Detection Using Clinical Images: A Comprehensive Review. *Life* **13**, 2123 (2023).
7. Grignaffini, F. *et al.* Machine Learning Approaches for Skin Cancer Classification from Dermoscopic Images: A Systematic Review. *Algorithms* **15**, 438 (2022).
8. Xia, M. *et al.* Lesion identification and malignancy prediction from clinical dermatological images. *Sci. Rep.* **12**, 15836 (2022).
9. Mundada, M. R., Seema, S., J, S. B. & Student, S. F. DeepDerm: Elevating Precision in Dermatological Diagnosis with Enhanced CNN. in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* 1–7 (2024). doi:10.1109/ICCCNT61001.2024.10724699.
10. Han, S. S. *et al.* Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *J. Invest. Dermatol.* **138**, 1529–1538 (2018).

11. Salido, J. A. & Ruiz, C. Using deep learning for melanoma detection in dermoscopy images. *Int. J. Mach. Learn. Comput.* **8**, 61–68 (2018).
12. Fujisawa, Y. *et al.* Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *Br. J. Dermatol.* **180**, 373–381 (2019).
13. Tschandl, P. *et al.* Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. *JAMA Dermatol.* **155**, 58–65 (2019).
14. Khan, S. & Khan, A. SkinViT: A transformer based method for Melanoma and Nonmelanoma classification. *PLOS ONE* **18**, e0295151 (2023).
15. Aksoy, S., Demircioglu, P. & Bogrekci, I. Enhancing Melanoma Diagnosis with Advanced Deep Learning Models Focusing on Vision Transformer, Swin Transformer, and ConvNeXt. *Dermatopathology* **11**, 239–252 (2024).
16. Wang, R. *et al.* A novel approach for melanoma detection utilizing GAN synthesis and vision transformer. *Comput. Biol. Med.* **176**, 108572 (2024).
17. Fang, A., Kornblith, S. & Schmidt, L. Does progress on ImageNet transfer to real-world datasets? *Adv. Neural Inf. Process. Syst.* **36**, 25050–25080 (2023).
18. Doerrich, S., Di Salvo, F., Brockmann, J. & Ledig, C. Rethinking model prototyping through the MedMNIST+ dataset collection. *Sci. Rep.* **15**, 7669 (2025).
19. Combalia, M. *et al.* Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. *Lancet Digit. Health* **4**, e330–e339 (2022).
20. Wolf, J. A. *et al.* Diagnostic Inaccuracy of Smartphone Applications for Melanoma Detection. *JAMA Dermatol.* **149**, 422–426 (2013).
21. Wongvibulsin, S. *et al.* Current State of Dermatology Mobile Applications With Artificial Intelligence Features. *JAMA Dermatol.* **160**, 646–650 (2024).

22. Sun, M. D. *et al.* Accuracy of commercially available smartphone applications for the detection of melanoma. *Br. J. Dermatol.* **186**, 744–746 (2022).
23. Wang, S. Q. *et al.* Differences in Dermoscopic Images from Nonpolarized Dermoscope and Polarized Dermoscope Influence the Diagnostic Accuracy and Confidence Level: A Pilot Study. *Dermatol. Surg.* **34**, 1389 (2008).
24. Hanlon, K. L., Wei, G., Correa-Selm, L. & Grichnik, J. M. Dermoscopy and skin imaging light sources: a comparison and review of spectral power distribution and color consistency. *J. Biomed. Opt.* **27**, 080902 (2022).
25. Petrie, T. C. *et al.* Quantifying Acceptable Artefact Ranges for Dermatologic Classification Algorithms. *Skin Health Dis.* **1**, ski2.19 (2021).
26. Daneshjou, R. *et al.* Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.* **8**, eabq6147 (2022).
27. Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V. & Zou, J. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. *JAMA Dermatol.* **157**, 1362–1369 (2021).
28. Mehta, D. *et al.* Multi-task AI models in dermatology: Overcoming critical clinical translation challenges for enhanced skin lesion diagnosis. *J. Eur. Acad. Dermatol. Venereol.* **n/a**,.
29. Kim, S., Gaibor, E., Matejek, B. & Haehn, D. Melanoma Detection with Uncertainty Quantification. Preprint at https://doi.org/10.48550/arXiv.2411.10322 (2024).
30. Shamsi, A. *et al.* A novel uncertainty-aware deep learning technique with an application on skin cancer diagnosis. *Neural Comput. Appl.* **35**, 22179–22188 (2023).
31. Ovadia, Y. *et al.* Can you trust your model' s uncertainty? Evaluating predictive uncertainty under dataset shift. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).

32. Schreyer, W. M., Anderson, C. & Thompson, R. F. Generalization is not a universal guarantee: Estimating similarity to training data with an ensemble out-of-distribution metric. Preprint at https://doi.org/10.48550/arXiv.2502.16329 (2025).
33. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 (2018).
34. Ricci Lara, M. A. *et al.* A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population. *Sci. Data* **10**, 712 (2023).
35. Pacheco, A. G. C. *et al.* PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data Brief* **32**, 106221 (2020).
36. Fitzpatrick, T. B. The validity and practicality of sun-reactive skin types I through VI. *Arch. Dermatol.* **124**, 869–871 (1988).
37. Mitchell, M. *et al.* Model Cards for Model Reporting. in *Proceedings of the Conference on Fairness, Accountability, and Transparency* 220–229 (Association for Computing Machinery, New York, NY, USA, 2019). doi:10.1145/3287560.3287596.
38. Howard, J. J., Sirotin, Y. B., Tipton, J. L. & Vemury, A. R. Reliability and Validity of Image-Based and Self-Reported Skin Phenotype Metrics. *IEEE Trans. Biom. Behav. Identity Sci.* **3**, 550–560 (2021).
39. Weir, V. R., Dempsey, K., Gichoya, J. W., Rotemberg, V. & Wong, A.-K. I. A survey of skin tone assessment in prospective research. *Npj Digit. Med.* **7**, 191 (2024).
40. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. Preprint at https://doi.org/10.48550/arXiv.1912.01703 (2019).
41. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009). doi:10.1109/CVPR.2009.5206848.

42. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. Preprint at https://doi.org/10.48550/arXiv.1711.05101 (2019).

43. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

44. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

45. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).