



MILK10k: A Hierarchical Multimodal Imaging-Learning Toolkit for Diagnosing Pigmented and Nonpigmented Skin Cancer and its Simulators

Philipp Tschandl^{1,2}, Bengu Nisa Akay³, Cliff Rosendahl⁴, Veronica Rotemberg⁵, Verche Todorovska⁶, Jochen Weber⁵, Anna Katharina Wolber⁷, Christoph Müller⁷, Nicholas Kurtansky⁵, Allan Halpern⁵, Wolfgang Weninger⁷ and Harald Kittler^{1,2}

Journal of Investigative Dermatology (2026) **146**, 357–364; doi:10.1016/j.jid.2025.06.1594

In this study, we introduce MILK10k, a multimodal image dataset designed to enhance machine learning and artificial intelligence–driven applications for diagnosing skin lesions suspected to be malignant neoplasms. MILK10k includes both close-up and dermatoscopy images expanding upon existing datasets with several enhancements. It expands disease coverage to 48 International Skin Imaging Collaboration–Designated Diagnoses—matched diagnoses, encompassing both neoplastic and non-neoplastic as well as pigmented and nonpigmented skin lesions. In addition, MILK10k incorporates images from individuals with diverse skin tones as well as rare skin diseases. The dataset includes 10,480 images from 5240 cases, retrospectively collected across 5 different centers. Of these, 95.7% ($n = 5016$) have been biopsied or excised, with histopathology serving as the ground truth. Accompanying metadata includes information on age, sex, skin tone, anatomic site, and diagnosis with varying levels of granularity. In addition to the dataset, we provide results from a machine-learning pipeline that evaluates images on the basis of common, human-interpretable concepts such as pigmentation, ulceration, hair, and skin markings. Furthermore, we provide a benchmark test set of 948 multimodal images from the same sources and an online tool for assessing key metrics on this set, creating a platform for evaluating both machine-learning models and human reader studies.

Keywords: Classification, Dataset, Dermatoscopy, Neural network, Ontology

INTRODUCTION

The application of machine learning (ML) and artificial intelligence (AI) has led to significant advancements in the automated assessment of medical images. Given the need for large training datasets to develop robust neural networks, substantial efforts have been made to collect and publish extensive image datasets. At the time of this submission, the

International Skin Imaging Collaboration (ISIC) Archive contained over 500,000 publicly available images. These datasets have driven the development of AI-based systems that, particularly in image classification tasks, have demonstrated experimental capabilities comparable with those of dermatologists and have the potential to improve diagnostic accuracy and efficiency (Barata et al, 2023; Marchetti et al, 2023; Tschandl et al, 2020). However, challenges remain in ensuring generalizability, fairness, and robustness of AI models. Techniques such as unsupervised, semisupervised, and multimodal/multitask learning as well as out-of-distribution detection offer promising solutions. Whereas larger datasets support general AI training, out-of-distribution detection requires a broader spectrum of diagnoses, and multimodal or unsupervised learning demands large sets of matched multi-instance or multimodal data (Chen et al, 2020¹; Wang et al, 2025; Yap et al, 2018). Existing datasets, including the ISIC Challenge datasets (Hernández-Pérez et al, 2024; Kurtansky et al, 2024; Rotemberg et al, 2021; Tschandl et al, 2018) and others such as PH2 (Mendonça et al, 2013), MED-NODE (Giotis et al, 2015), PAD-UFES-20 (Pacheco et al, 2020), SD-198 (Sun et al, 2016), and

¹Division for Applied AI Research in Dermatology, Department of Dermatology, Medical University of Vienna, Vienna, Austria;

²Comprehensive Center for Artificial Intelligence in Medicine, Medical University of Vienna, Vienna, Austria; ³Department of Dermatology, Faculty of Medicine, Ankara University, Ankara, Turkey; ⁴Mayne Academy of General Practice, Medical School, The University of Queensland, Brisbane, Australia; ⁵Dermatology Service, Memorial Sloan Kettering Cancer Center, New York, New York, USA; ⁶Dermatology Private Practice, Skopje, North Macedonia; and ⁷Department of Dermatology, Medical University of Vienna, Vienna, Austria

Correspondence: Harald Kittler, Department of Dermatology, Medical University of Vienna, Währinger Gürtel 18-20, Vienna 1090, Austria. E-mail: harald.kittler@meduniwien.ac.at

Abbreviations: AI, artificial intelligence; BCC, basal cell carcinoma; CI, confidence interval; ISIC, International Skin Imaging Collaboration; ISIC-DX, International Skin Imaging Collaboration–Designated Diagnoses; KA, keratoacanthoma; ML, machine learning; SCC, squamous cell carcinoma

Received 14 April 2025; revised 18 June 2025; accepted 25 June 2025; accepted manuscript published online 21 July 2025; corrected proof published online 19 August 2025

¹ Chen T, Kornblith S, Swersky K, Norouzi M, Hinton G. Big self-supervised models are strong semi-supervised learners. arXiv 2020.

BALD (Müller et al, 2025), focus on a single imaging modality at a time. One legacy dataset (Kawahara et al, 2019) includes 2045 images of pigmented skin lesions, including both dermatoscopic and close-up images, but this dataset is limited mainly to pigmented lesions and covers only a limited range of diagnoses. Furthermore, most datasets remain limited in diversity with no information on skin tone, leading to potential biases in both datasets and resulting algorithms (Groh et al, 2022², 2021). Recent efforts address these limitations by crowdsourcing data collection (Ward et al, 2024), improving skin tone representation (Daneshjou et al, 2022), and expanding data collection to underrepresented geographic areas (Ricci Lara et al, 2023; Yilmaz et al, 2024). However, very few datasets align with a public diagnostic disease ontology, an issue that the recently proposed International Skin Imaging Collaboration-Designated Diagnoses (ISIC-DX) hierarchy (Scope et al, 2025) aims to address as a future standard for clinically meaningful disease classifications and model assessment (Mahmood, 2025; Yilmaz et al, 2024). The dataset presented in this paper seeks to fill these gaps by providing a large, multicenter, multimodal dataset of skin lesions, including nonpigmented lesions and a broader range of diagnostic classes that have been aligned to the ISIC-DX hierarchy. Beyond the dataset, we provide a robust ML pipeline and AI-powered data auditing and offer a benchmark test set.

MATERIALS AND METHODS

Ethics review

This dataset consists of retrospectively collected, deidentified images and pseudonymized metadata gathered during routine patient care at 3 tertiary referral centers (Department of Dermatology, Medical University of Vienna, Vienna, Austria; Department of Dermatology, Ankara University, Ankara, Turkey; and the Dermatology Service, Memorial Sloan Kettering Cancer Center, New York, NY), 1 dermatology practice (Skopje, North Macedonia), and 1 general medicine practice with special interest in skin cancer (Capalaba, Australia). Each participating center complied with its respective data protection and ethical regulations during image collection. At the Department of Dermatology, Medical University of Vienna, data collection was approved by the ethics board of the Medical University of Vienna (ethics protocol numbers 1081/2015, 1536/2017, and 1070/2023) and underwent an additional review by the institutional data-clearing office. Images from the general practice in Queensland, Australia were collected under ethics protocol numbers 2015000162 and 2017001223 (University of Queensland), and those from Memorial Sloan Kettering were obtained in accordance with their institutional review board protocol number 16-974. Data from the Department of Dermatology at the Ankara University, Turkey and from the dermatology practice in Skopje, North Macedonia were contributed to the European Dermoscopy Database, hosted by the Department of Dermatology at the Medical University of Vienna, under ethics protocol number 1536/2017 (ethics board of the Medical University of Vienna). No written informed consent was sought because only nonidentifiable retrospective data were collected.

² Groh M, Harris C, Daneshjou R, Badri O, Koochek A. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. arXiv 2022.

Data collection and selection

We included high-quality close-up and dermatoscopic images of pigmented lesions as well as nonpigmented lesions selected for biopsy or excision when skin cancer was considered in the differential diagnosis. For these lesions, histopathology served as the ground truth. To minimize verification bias, we also included a limited number of diagnostically straightforward lesions without histopathological confirmation, where clinical diagnosis was established by the treating physician and confirmed by consensus between 2 authors (PT and HK). Nail pigmentations were excluded.

Images were continuously collected from the Department of Dermatology in Vienna and the general practice in Queensland, whereas cases from other institutions were obtained through convenience sampling. To avoid substantial overlap with the HAM10000 dataset, which contains only dermatoscopic images of pigmented lesions, we excluded pigmented lesions from the Queensland general practice if they were already part of HAM10000. This approach also enriched our dataset with more challenging nonpigmented lesions, which remain underrepresented in publicly available dermatoscopic image sets for skin cancer diagnosis.

During an initial quality assessment, we excluded images of poor quality, those containing identifiable content (eg, tattoos), and cases with ambiguous pathology reports. Close-up images were captured using a variety of cameras, without standardized photography protocols across centers. At the Department of Dermatology in Vienna, images were taken by a professional photographer in a dedicated studio, whereas at other centers, they were captured by treating physicians or medical personnel. Dermoscopy images were acquired using polarized light.

Pathologic diagnoses were collected from the internal documentation databases of the centers. Diagnoses were manually mapped to the most specific ISIC-DX category possible as well as to the ISIC2018 terminology and HAM10000-level classification by 2 authors (PT and HK). Melanoma invasion thickness was categorized into the following intervals: <0.8 mm, ≥0.8 to 1 mm, ≥1 to 2 mm, ≥2 to 3 mm, ≥3 to 4 mm, and ≥4 mm, which correspond to T stages T1a, T1b, T2, T3, and T4 of melanoma staging of International Union Against Cancer/American Joint Committee on Cancer 8th edition, respectively. Patient age was rounded to the nearest 5-year increment, and anatomic locations were manually assigned to one of the following categories: trunk, head/neck/face, lower extremity, upper extremity, hand, foot, or genital.

Data processing

All images were subject to removal of EXIF metadata, normalization of aspect ratio, landscape orientation, file format, and resolution as described previously (Tschandl et al, 2018). To avoid potential reidentification of subjects, visible anatomic landmarks or dermatoglyphs surrounding the lesion were cropped and/or blurred. Any suspected duplicates or labeling errors related to diagnosis, image type, size, or skin tone, detected using SelfClean (Gröger et al, 2025) were removed. Images were annotated with the MONET framework (Kim et al, 2024) with the following concept term groups and their probability added to the metadata: (i) "ulceration" and "crust"; (ii) "hair"; (iii) "vasculature" and "vessels"; (iv) "erythema"; (v) "pigmented"; (vi) "gel," "water drop," "fluid," and "dermoscopy liquid"; and (vii) "skin markings," "pen ink," and "purple pen."

Owing to the pseudonymization process, no fixed mapping remained between previously published images and MILK10k cases.

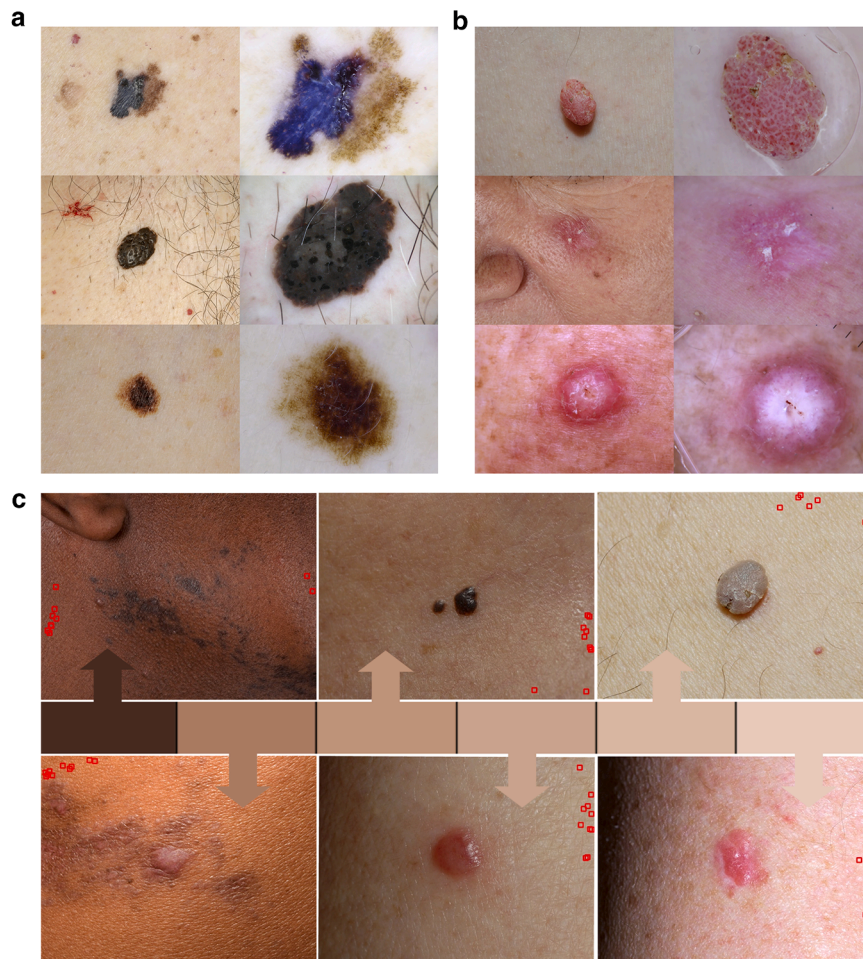


Figure 1. Corresponding image pairs and skin tone extraction from close-up images. (a) Three examples of corresponding close-up and dermatoscopic images of pigmented lesions. (b) Three examples of corresponding close-up and dermatoscopic images of nonpigmented lesions. (c) Skin tone is extracted from the periphery of close-up images by calculating the median color of ten 10×10 -pixel crops (red squares). On the basis of the range of extracted skin tones, we created 6 skin tone clusters sorted by lightness. In addition, an expert reviewed the skin tone assignment for each image and made corrections as needed.

To retrospectively link MILK10k cases to existing HAM10000 lesions, a stepwise combination of automated and manual matching was applied: first, we calculated image features of both MILK10k images and existing HAM10000 images with the transformers library using the pipeline(task='image-feature-extraction', model_name='facebook/dinov2-base') function (Oquab et al, 2023³). Image pairs were then compared on the basis of cosine similarity and manually reviewed to determine cutoff thresholds for classification: similarities above 0.93 were deemed "certain matches," those between 0.80 and 0.93 were deemed as "possible matches," and those below 0.80 were deemed as "certain nonmatches." For image pairs in the intermediate range (0.80–0.93), the top 6 most similar candidates per lesion were manually reviewed. Through this process, 2196 lesions were automatically discarded as nonmatches, and 2059 were manually discarded. In contrast, 392 lesions were automatically matched, and 593 were manually confirmed as matches with existing HAM10000 lesions. These identified matches were integrated into the dataset metadata. Although this approach is heuristic and may include some degree of approximation, it provides a practical and reasonably accurate method for identifying duplicate or overlapping cases.

Skin tone

To automatically extract skin tone from close-up images, we generated 500 random crops (10×10 pixels) from the periphery of each image. We filtered these crops on the basis of gradient, brightness, and color dominance to exclude those that included the standard blue or black background walls commonly used in clinical photography. To avoid bias toward darker skin tones due to the inclusion of hair, we sorted the valid crops by lightness and selected the top 10. The median color of these top 10 crops represented the extracted skin tone. We then used K-Means clustering to create 6 skin tone cluster exemplars, which we sorted by their lightness (L^* component) after converting them from RGB to Lab color space (Figure 1c). All automated skin tone classifications were manually checked and revised if necessary. Reclassification was particularly necessary for close-ups of the scalp and face, where the automatic extraction of skin tone was occasionally incorrect owing to hair or severe chronic sun damage of the surrounding skin. The resulting L -values (mean $\pm$ SD) for tone clusters 0–5 were 52.1 ± 13.7 , 53.7 ± 6.1 , 61.9 ± 4.6 , 66.7 ± 6.1 , 75.7 ± 7.3 , and 83.0 ± 2.9 , respectively.

Dataset description

MILK10K consists of 10,480 images corresponding to image pairs of 5240 lesions, featuring 1 clinical close-up image and its corresponding dermatoscopic image. The anatomical site distribution is illustrated in Table 1, with most lesions found on the trunk and in the

³ Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: learning robust visual features without supervision arXiv 2023.

Table 1. Main Characteristics of 5240 Lesions in the MILK10K Dataset		
Characteristics	n	%
Anatomic site		
Trunk	1925	36.7
Head and neck	1491	28.5
Lower extremity	1023	19.5
Upper extremity	667	12.7
Hand	72	1.4
Foot	23	0.4
Genital	8	0.2
missing	31	0.6
Skin tone		
0.0 (darkest)	6	0.1
1.0	105	2.0
2.0	506	9.7
3.0	3174	60.6
4.0	1066	20.3
5.0 (lightest)	383	7.3
Diagnostic class		
Basal cell carcinoma	2522	48.1
Melanocytic nevus, any type	746	14.2
Benign keratinocytic lesion	544	10.4
Squamous cell carcinoma/keratoacanthoma	473	9.0
Melanoma	450	8.6
Actinic keratosis/intraepidermal carcinoma	303	5.8
Dermatofibroma	52	1.0
Inflammatory and infectious	50	1.0
Vascular lesions and hemorrhage	47	0.9
Other benign proliferations including collisions	44	0.8
Other malignant proliferations including collisions	9	0.2

head and neck areas. The mean age of the patients was 61.1 (SD = 16.1) years, with 39.8% (n = 2088) being female. Skin tone was categorized into 6 levels, ranging from very dark to very light, as shown in Figure 1. To prevent confusion with the Fitzpatrick skin types, the skin tone categories were arranged so that 0 represents the darkest skin tone, and 5 represents the lightest. Most patients had skin tones in the middle ranges (Table 1). Of the 5240 lesions, 5016 (95.7%) were biopsied or excised, where histopathology served as the gold standard. The histopathologic diagnoses were mapped to the ISIC-DX diagnostic scheme as well as a simplified scheme on the basis of the ISIC2018 and HAM10000 diagnostic categories (Table 2). The number of included granular ISIC-DX classes is shown in Supplementary Table S1 and Supplementary Figure S1.

To improve the interpretability of AI-driven ML experiments and uncover potential biases, we annotated the images with human-interpretable concepts using the MONET framework (Kim et al, 2024). We included concepts known to correlate with specific diagnoses, such as pigmentation, ulceration or crust, vessels, and erythema. In addition, we incorporated concepts that might indicate spurious correlations or artifacts, such as hair, contact fluid, and skin markings. Figure 2 presents the top 10 examples for each concept, whereas histograms illustrating the distribution of concept values across diagnostic categories are provided in Supplementary Figure S2. To validate the annotations, we examined selected concepts in more detail and assessed their plausibility. As expected, significantly higher pigmentation values were observed in

melanomas and nevi than in other diagnostic classes (Supplementary Figure S2). In contrast, erythema values were highest in inflammatory lesions and basal cell carcinomas (BCCs). Similarly, the presence of ulceration or crust was most pronounced in squamous cell carcinoma (SCC)/keratoacanthoma (KA). The presence of vessels and visible vasculature was most prominent in vascular lesions and in BCCs but showed little distinction from inflammatory conditions or diagnoses associated with sun-damaged skin. This indicates that MONET analyses are not limited to the lesion area alone and demonstrates that the mere presence of vasculature in an image is not predictive of any specific diagnosis.

In addition to the MILK10K dataset, we provide a basic ML pipeline to demonstrate potential handling of a multiclass classification using both a closeup and dermatoscopic image with hierarchical labels (details are provided in Data Availability Statement). Using fivefold cross-validation with 11 generic diagnostic classes as target categories, our pipeline achieved an average recall score of 0.426 (95% confidence interval [CI] = 0.392–0.459) (Figure 3 and Supplementary Figure S3), with a corresponding mean specificity of 0.960 (95% CI = 0.959–0.961). This relatively lower performance than existing ISIC challenges suggests that the inclusion of non-pigmented cases and challenging classes (eg, inflammatory and collision lesions) increases the classification difficulty (Combalia et al, 2022; Tschandl et al, 2019).

When using the granular ISIC-DX (48-class) diagnostic labels, we obtained a mean top-1 accuracy of 53.6% (95% CI = 52.5–54.6) and a mean top 3 accuracy of 67.7% (95% CI = 66.2–69.2). Traditional performance metrics, such as these, may be less informative when dealing with large sets of diagnostic categories across multiple levels of granularity. For instance, misclassifying "nevus, compound" as "nevus" is penalized equally as misclassifying it as "SCC, invasive," despite the former being a much closer diagnosis.

To address this limitation, we propose a metric based on an underlying hierarchical terminology such as ISIC-DX (Supplementary Figure S4): the average distance between predicted and true diagnoses, measured as steps within a structured classification hierarchy. This metric can theoretically range from 0 to the longest shortest path length between any 2 nodes of the hierarchy. To prevent inconsistencies in a purely linear hierarchy, we introduced a "root" node connecting the least granular categories (malignant, indeterminate, benign). We anticipate that future community-driven refinements, such as creating fully connected hierarchies, considering management overlaps, transformations to ranking tasks, reweighting relevant differential diagnoses, or simply normalization of the metric, will enhance the utility of this raw distance-based calculation. Using our pipeline, we obtained an average prediction-to-ground-truth distance of 1.73 (95% CI = 1.66–1.79).

Eventually, Supplementary Figure S5 provides a schematic overview of described workflows steps while processing the dataset.

A separate benchmark dataset with undisclosed diagnoses

Because independent test sets are often required for ML experiments and accompanying reader studies, we have curated a smaller benchmark dataset alongside MILK10k. This smaller dataset consists of 958 images, corresponding to pairs of close-up and dermatoscopic images for 479 lesions. It is derived from the same sources as MILK10k and covers the same diagnostic categories. The dataset was constructed from a master set using a stratified approach (stratified by diagnostic categories) with additional constraints, which remain

Table 2. Mapping of ISIC-DX Diagnostic Classes to a Simplified schema

Diagnostic Class	Abbreviation	ISIC-DX
Actinic keratosis/intraepidermal carcinoma	AK/IEC	Solar or actinic keratosis, squamous cell carcinoma in situ, and Bowen's disease
Basal cell carcinoma	BCC	Basal cell carcinoma
Other benign proliferations including collisions	BEN_OTH	Benign - other, benign soft tissue proliferations - Fibrohistiocytic, benign soft tissue proliferations - vascular, collision - only benign proliferations, cylindroma, exogenous, fibroepithelial polyp, fibroma, infundibular or epidermal cyst, juvenile xanthogranuloma, mastocytosis, mucosal melanotic macule, scar, sebaceous hyperplasia, spiradenoma, supernumerary nipple, trichilemmal or isthmic-catagen or pilar cyst, trichoblastoma
Benign keratinocytic lesion	BKL	Clear cell acanthoma, ink-spot lentigo, lichen planus-like keratosis, seborrheic keratosis, and solar lentigo
Dermatofibroma	DF	Dermatofibroma
Inflammatory and infectious	INF	Inflammatory or infectious diseases, molluscum, porokeratosis, verruca
Other malignant proliferations, including collisions	MAL_OTH	Atypical fibroxanthoma, collision—at least 1 malignant proliferation, Kaposi sarcoma, lymphocytic proliferations - T-cell/NK, malignant peripheral nerve sheath tumor, Merkel cell carcinoma
Melanoma	MEL	Melanoma invasive, melanoma in situ, and melanoma metastasis
Melanocytic nevus, any type	NV	Blue nevus, nevus, nevus, acral, nevus, BAP-1 deficient, nevus, balloon cell, nevus, combined, nevus, congenital, nevus, deep penetrating, nevus, NOS, compound, nevus, NOS, dermal, nevus, NOS, junctional, nevus, recurrent or persistent, nevus, Reed, nevus, Spilus, nevus, Spitz
Squamous cell carcinoma/keratoacanthoma	SCC/KA	Keratoacanthoma and squamous cell carcinoma, invasive
Vascular lesions and hemorrhage	VASC	Angiokeratoma, arterio-venous malformation, hemangioma, hemangioma, hobnail, lymphangioma, and pyogenic granuloma

Abbreviation: ISIC-DX, International Skin Imaging Collaboration-Designated Diagnoses. Notably, the ISIC-DX "Benign - Other" class is a subset of the larger "BEN_OTH" category.

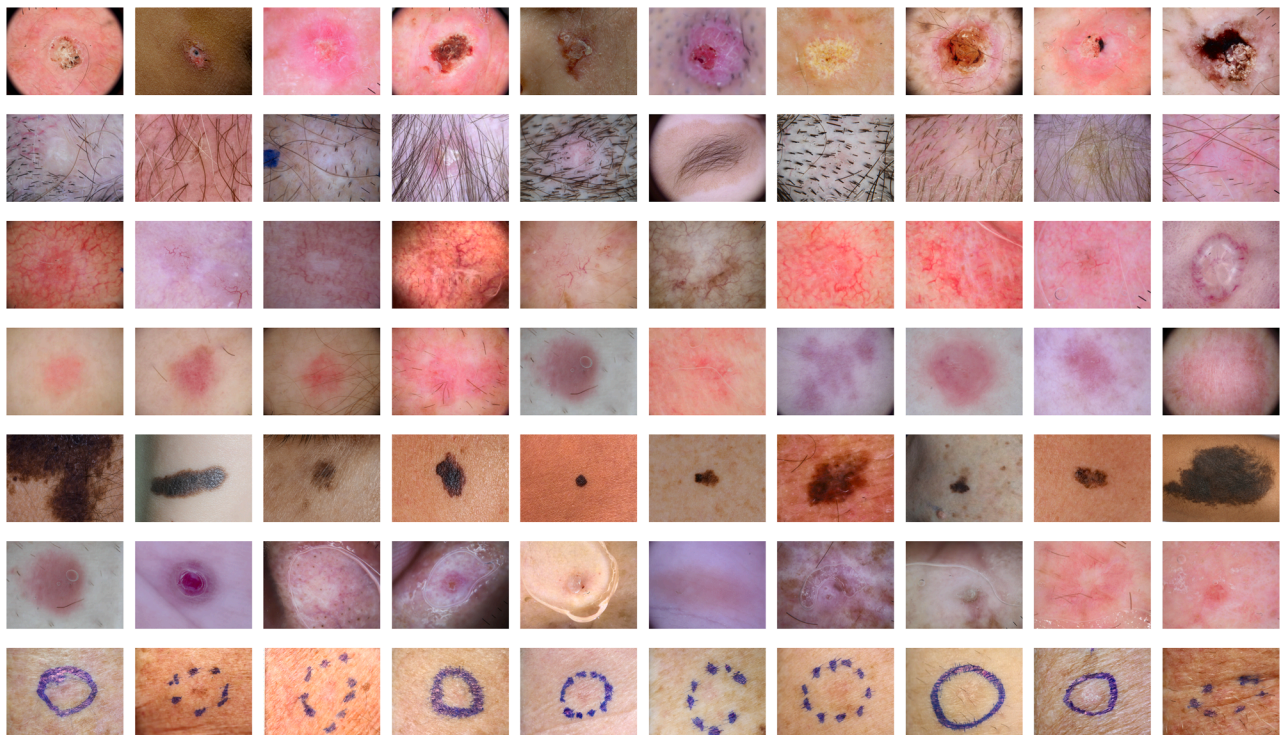


Figure 2. Images with top 10 values for specific visual attributes as obtained through the MONET analysis pipeline. Rows from top to bottom show (1) "ulceration" and "crust"; (2) "hair"; (3) "vasculature" and "vessels"; (4) "erythema"; (5) "pigmented"; (6) "gel," "water drop," "fluid," and "dermoscopy liquid"; and (7) "skin markings," "pen ink," and "purple pen."

undisclosed to prevent ground truth leakage. Although the diagnostic categories align with those in MILK10k, the specific diagnoses may differ, simulating a domain shift. The benchmark test set

includes additional diagnoses not present in MILK10k, whereas some diagnoses found in MILK10k are absent. This primarily applies to the 2 categories: "other benign" and "other malignant

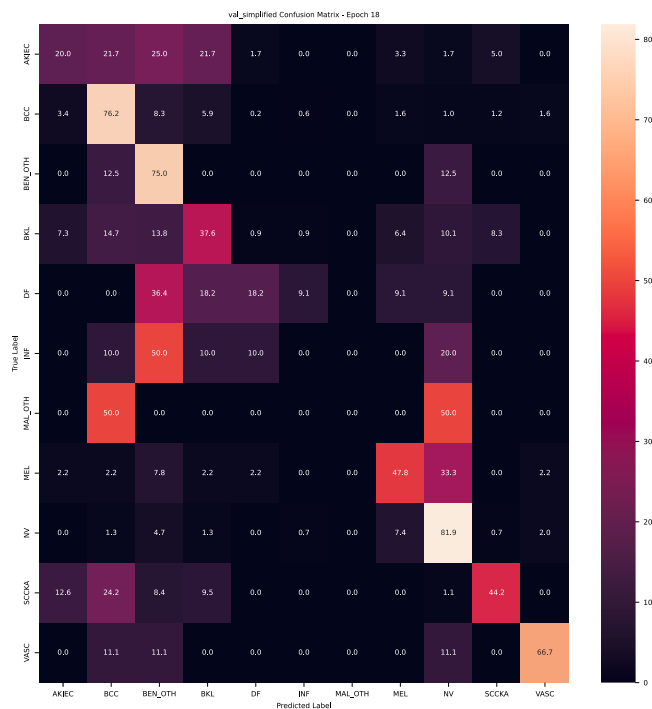


Figure 3. Confusion matrix of 1 cross-validation run of classification fine tuning. Numbers denote fractions normalized by the true class (rows). Although it seems that none of the MAL_OTH class was predicted correctly, this could still be a reasonable result if classes truly depict a collision where one of the contained entities is a BCC. AKIEC, actinic keratosis intraepidermal carcinoma; BCC, basal cell carcinoma; BKL, benign keratinocytic lesion; DF, dermatofibroma; INF, inflammatory and infectious; MEL, melanoma; NV, nevus; SCCA, squamous cell carcinoma/keratoacanthoma; VASC, vascular lesions and hemorrhage.

proliferations." This design enhances the benchmark dataset's ability to assess robustness to domain shifts and supports its use for zero-shot learning. Although we do not disclose specific diagnoses or diagnostic categories, all possible granular ISIC-DX classes are included in the mapping in Table 2. The benchmark dataset is available as part of a live challenge within the ISIC framework and can be downloaded from the ISIC archive (doi: <https://doi.org/10.34970/648456>). To preserve its role as an independent external benchmark for ML experiments and reader studies, diagnoses for this set will remain undisclosed for the next 5 years. The online tool for assessing key metrics is implemented as an ISIC live challenge and can be accessed at <https://challenge.isic-archive.com/landing/milk10k/>.

Usage notes

It is important to note that the skin tone measurements reflect facultative skin tone, which can be altered by sun exposure, rather than an individual's unaltered constitutional skin tone. In addition, individuals with darker skin tones remain underrepresented in the dataset. One reason for this is that cases were primarily included when a biopsy or excision was performed to confirm or exclude suspected skin cancer, which is more common in individuals with lighter skin tones. As a result, the dataset is mainly useful for diagnosing skin cancer in individuals with light-to-medium skin tones. Children are also underrepresented because skin cancer is rare in pediatric populations (Paulson et al, 2020), limiting the applicability of any derived algorithm for diagnosing lesions in children.

Furthermore, the dataset does not include nail pigmentations, and certain anatomical sites such as the mucosa, genital areas, and acral regions are underrepresented, which affects generalizability. Because the vast majority of lesions were biopsied or excised, the dataset is also subject to verification bias, meaning that benign, clinically unremarkable lesions are underrepresented. It is important to note that the prevalence of malignant diseases in this dataset does not reflect the typical distribution seen in clinical practice in the general population. As a result, model performance on this dataset may not accurately represent its effectiveness in a real-world medical screening context. Histopathologic diagnoses must be regarded as a "best effort" because different sampling techniques or failure to perform deeper cuts may miss additional lesions or more invasive parts of a tumor. Users should acknowledge diagnostic limitations in some cases. For example, irritated seborrheic keratoses may be indistinguishable from irritated viral warts, nodular melanoma may be indistinguishable from melanoma metastases, and some pathologists do not differentiate Reed nevi from Spitz nevi. Users should also be aware that some diagnoses are not mutually exclusive. For example, an acral nevus can also be congenital, and a junctional nevus can be "atypical." There is no consensus until date on how to resolve this situation; for example, the DERM12345 dataset (Yilmaz et al, 2024), having a partly overlapping taxonomy to ISIC-DX, assigned multiple main and subclasses to a case.

Collision lesions are classified under "other benign proliferations" if both lesions are benign and under "other malignant proliferations" if at least 1 lesion is malignant. Melanomas that develop within a nevus are not classified as collision lesions because both melanomas and nevi originate from the same cell type, namely melanocytes. A collision lesion is defined as 2 proliferations originating from different cell types. If a BCC collides with a nevus, it is categorized as "other malignant proliferations" rather than BCC alone. This means that even if a model correctly predicts BCC, the prediction would be considered incorrect because the correct label for this collision is "other malignant proliferations." For ML experiments, we generally recommend first predicting the most specific diagnosis and then using this prediction to assign cases to the 11 generic categories for metrics that rely on category labels as the primary outcome variable. Although all cases labeled as "other malignant proliferations" in the current dataset are collision tumors, future efforts should extend beyond this because this category may also include a variety of other, often rare tumor types. For nonmedical users of the dataset, we further recommend studying the description of the generic categories given in the HAM10000 paper. The introduced categories—"inflammatory and infectious lesions," "other benign proliferations," and "other malignant proliferations"—encompass a diverse mix of diseases that may be only distantly related morphologically. As a result, an image randomly selected from 1 of these classes may, on average, be more dissimilar to another image within the same class than to an image from a different class.

Cutaneous B-cell lymphomas are mapped to the category "malignant other proliferations," even though some underlying subclasses may follow a indolent, self-limiting course. Mastocytosis is labeled as "benign other proliferations," although the disease may have a malignant course in adult patients.

Another diagnostic category introduced in this dataset, which was not included in HAM10000, is SCC/KA. Owing to the retrospective nature of the dataset, we cannot reliably distinguish between KA and highly differentiated SCC. There is ongoing debate about whether KA represents a benign, intermediate, or malignant keratinocytic

proliferation. In many cases, especially in partial biopsies, it is difficult to differentiate between KA and SCC pathologically. For this reason, they have been grouped together, which aligns with the 5th edition of the [World Health Organization Classification of Skin Tumors \(2023\)](#), where both are classified as malignant keratinocytic tumors. However, the ISIC-DX diagnostic tree contains both "KA" as part of "malignant epidermal proliferations" and "KA-like SCC" as part of "SCC, invasive." In accordance with the World Health Organization classification, we do not use the latter term, resulting in an empty class. This distinction may require reassessment by dataset users, particularly when integrating future datasets where classification schemes may differ.

The generic category of actinic keratosis/intraepidermal carcinoma groups together 2 types of noninvasive keratinocytic proliferations. Whereas some authors consider them precursors to SCC, others classify them as in situ carcinomas.

Bowen's disease is regarded as a synonym for intraepidermal carcinoma and is also included in this category. It is important to note that the 2019 ISIC Challenge differentiated between actinic keratosis and intraepidermal carcinoma, with a distinct category specifically for actinic keratosis only.

In general, the hierarchical structure of skin lesion classes is certainly more complex for users than a flat set of labels but could enable more precise analysis for management decisions and facilitate the merging of datasets with varying granularities. Furthermore, it could aid in identifying missing subtypes and refining the definition of relevant differential diagnoses, potentially enhancing the overall utility of future AI systems.

For privacy reasons, we have not disclosed the sources of individual images and strongly discourage any attempts to reverse engineer this information. The MILK10k dataset and the smaller benchmark dataset are provided under a CC BY-NC 4.0 license, meaning that they are intended strictly for teaching and research purposes. Any other use, including commercial applications, is prohibited. If users are uncertain that their intended use falls within the scope of this license, they should contact the corresponding author for clarification.

ETHICS STATEMENT

The source data were not created in this study, and data are regarded non-identifiable; thus, no statement was required. "Ethics review" in Materials and Methods provides more details.

DATA AVAILABILITY STATEMENT

Images and metadata are available at the ISIC-Archive as a pinned dataset (doi: <https://doi.org/10.34970/648456>), as well as the Harvard Dataverse (<https://doi.org/10.7910/DVN/FSXRAQ>). The benchmark test set is available at the ISIC Archive (<https://challenge.isic-archive.com/data/#milk10k>, <https://doi.org/10.34970/262082>), with ground-truth data available publicly after 5 years. Code for a basic machine learning pipeline and hierarchical ISIC-DX evaluation is available at Codeberg (https://codeberg.org/ptschandl/MILK10k_train_base). The MONET concept annotation pipeline is available at <https://github.com/suinleelab/MONET>.

CONFLICT OF INTEREST

PT reports speaker fees from AbbVie and unrestricted grants for education projects from Lilly. HK reports speaker fees from Fotofinder, MSD, Novartis, and Pelpharma; license fees from Casio; equipment from Fotofinder, Heine, and Casio; and advisor for Fotofinder, La Roche-Posay, and AI Medical Technology, member of the executive board of the International Dermoscopy Society. The remaining authors state no conflict of interest.

ACKNOWLEDGMENTS

We are grateful for the photographer of the Department of Dermatology at the Medical University of Vienna, Andreas Ebner, without whom the data collection at this institution would not be possible. Claude 3.5 Sonnet

(Anthropic PBC, San Francisco, CA) and Le Chat (Mistral AI SAS, Paris, France) were used to draft code snippets for data analysis, cleaning, and machine learning. No external funding sources or institutional funds were used to cover the research, authors' time, administrative support, and publishing costs.

AUTHOR CONTRIBUTIONS

Conceptualization: PT, HK; Data Curation: PT, AKW, HK; Formal Analysis: PT, HK; Investigation: PT, HK; Methodology: PT; Project Administration: HK; Resources: PT, BNA, AH, CM, CR, VR, VT, JW, WW, HK; Software: PT, JW, HK; Supervision: HK; Validation: PT, HK; Visualization: PT, HK; Writing - Original Draft Preparation: PT, HK; Writing - Review and Editing: PT, BNA, AH, CM, CR, VR, VT, JW, WW, AKW, HK

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at www.jidonline.org, and at [10.1016/j.jid.2025.06.1594](https://doi.org/10.1016/j.jid.2025.06.1594).

REFERENCES

- Barata C, Rotemberg V, Codella NCF, Tschandl P, Rinner C, Akay BN, et al. A reinforcement learning model for AI-based decision support in skin cancer. *Nat Med* 2023;29:1941–6.
- Combalia M, Codella N, Rotemberg V, Carrera C, Dusza S, Gutman D, 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* 2022;4:E330–9.
- Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv* 2022;8:eabq6147.
- Giotis I, Molders N, Land S, Biehl M, Jonkman MF, Petkov N. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Syst Appl* 2015;42:6578–85.
- Gröger F, Lionetti S, Gottfrois P, Gonzalez-Jimenez A, Amruthalingam L, Consortium L, et al. Intrinsic self-supervision for data quality audits. In: Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, et al., editors. *NIPS '24: Proceedings of the 38th international conference on neural information processing systems*. Red Hook, New York: Curran Associates Inc.; 2025. p. 92273–316.
- Groh M, Harris C, Soenksen L, Lau F, Han R, Kim A, et al. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. https://openaccess.thecvf.com/content/CVPR2021W/ISIC/papers/Groh_Evaluating_Deep_Neural_Networks_Trained_on_Clinical_Images_in_Dermatology_CVPRW_2021_paper.pdf; 2021 (accessed August 8, 2025).
- Hernández-Pérez C, Combalia M, Podlipnik S, Codella NCF, Rotemberg V, Halpern AC, et al. BCN20000: dermoscopic lesions in the wild. *Sci Data*. Springer Science and Business Media LLC 2024;11:641.
- Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE J Biomed Health Inform* 2019;23:538–46.
- Kim C, Gadgil SU, DeGrave AJ, Omiye JA, Cai ZR, Daneshjou R, et al. Transparent medical image AI via an image-text foundation model grounded in medical literature. *Nat Med* 2024;30:1154–65.
- Kurtansky NR, D'Alessandro BM, Gillis MC, Betz-Stablein B, Cerminara SE, Garcia R, et al. The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection. *Sci Data* 2024;11:884.
- Mahmood F. A benchmarking crisis in biomedical machine learning. *Nat Med* 2025;31:1060.
- Marchetti MA, Cowen EA, Kurtansky NR, Weber J, Dauscher M, DeFazio J, et al. Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study). *NPJ Digit Med* 2023;6:127.
- Mendonça T, Ferreira PM, Marques JS, Marcal AR, Rozeira J. PH² - a dermoscopic image database for research and benchmarking. *Annu Int Conf IEEE Eng Med Biol Soc* 2013;2013:5437–40.
- Müller C, Tschandl P, Rinner C, Kyrgidis A, Koga H, Moscarella E, et al. The BRAAFF-annotated acral lesions dataset (BALD): a curated set of dermoscopic images of acral melanoma and nevi from various sources. *J Invest Dermatol* 2025;145:1802–5.
- Pacheco AGC, Lima GR, Salomão AS, Krohling B, Biral IP, de Angelo GG, et al. PAD-UFES-20: a skin lesion dataset composed of

- patient data and clinical images collected from smartphones. *Data Brief* 2020;32:106221.
- Paulson KG, Gupta D, Kim TS, Veatch JR, Byrd DR, Bhatia S, et al. Age-specific incidence of melanoma in the United States. *JAMA Dermatol* 2020;156:57–64.
- Ricci Lara MA, Rodríguez Kowalczyk MV, Lisa Eliceche M, Ferrareso MG, Luna DR, Benitez SE, et al. A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population. *Sci Data* 2023;10:712.
- Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, Chousakos E, Codella N, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci Data* 2021;8:34.
- Scope A, Liopyris K, Weber J, Barnhill RL, Braun RP, Curiel-Lewandrowski CN, et al. International Skin Imaging Collaboration-Designated Diagnoses (ISIC-DX): consensus terminology for lesion diagnostic labeling. *J Eur Acad Dermatol Venereol* 2025;39:117–25.
- Sun X, Yang J, Sun M, Wang K. A benchmark for automatic visual classification of clinical skin disease images. *Computer vision – ECCV 2016*. In: Leibe B, Matas J, Sebe N, Welling M, editors. *Computer vision – ECCV 2016. Lecture notes in computer science*, vol 9910. Cham, Switzerland: Springer International Publishing; 2016. p. 206–22.
- Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. *Nat Med* 2020;26:1229–34.
- Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, Braun RP, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol* 2019;155:58–65.
- Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data* 2018;5:180161.
- Wang J, Zhang Y, Xie F, Liu J. Enhancing diagnosis of psoriasis and inflammatory skin diseases: a spatially aligned multimodal model integrating clinical and dermoscopic images. *J Invest Dermatol* 2025. S0022-202X(25)00402-00406.
- Ward A, Li J, Wang J, Lakshminarasimhan S, Carrick A, Campana B, et al. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Netw Open* 2024;7:e2446615.
- WHO Classification of Tumours: skin tumours [Internet; beta version ahead of print] [Internet]. Lyon, France: International Agency for Research on Cancer; 2023. <https://tumourclassification.iarc.who.int/chapters/64>. (accessed March 31, 2025).
- Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. *Exp Dermatol* 2018;27:1261–7.
- Yilmaz A, Yasar SP, Gencoglan G, Temelkuran B. DERM12345: a large, multisource dermatoscopic skin lesion dataset with 40 subclasses. *Sci Data* 2024;11:1302.



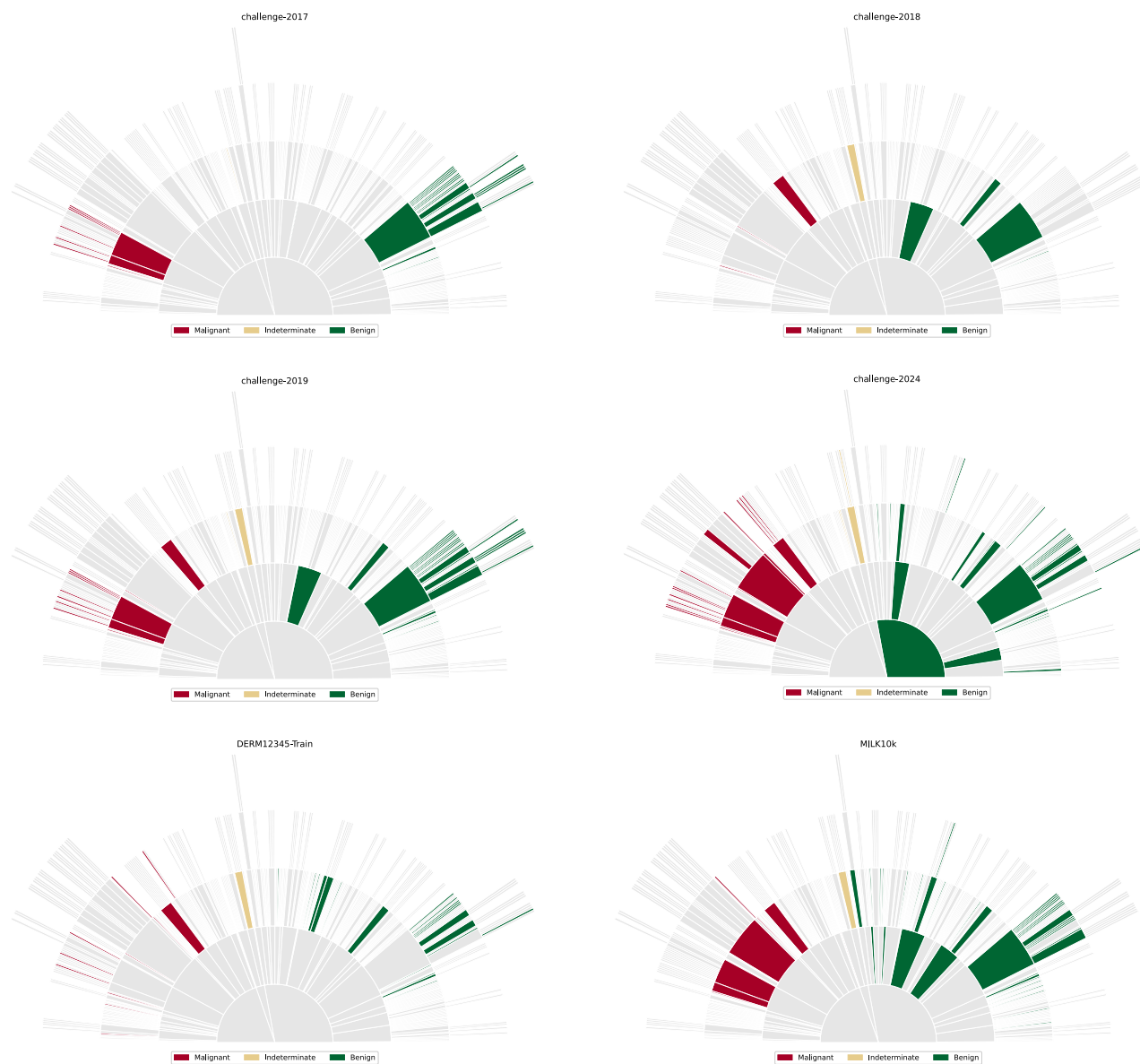
This work is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

SUPPLEMENTARY MATERIALS

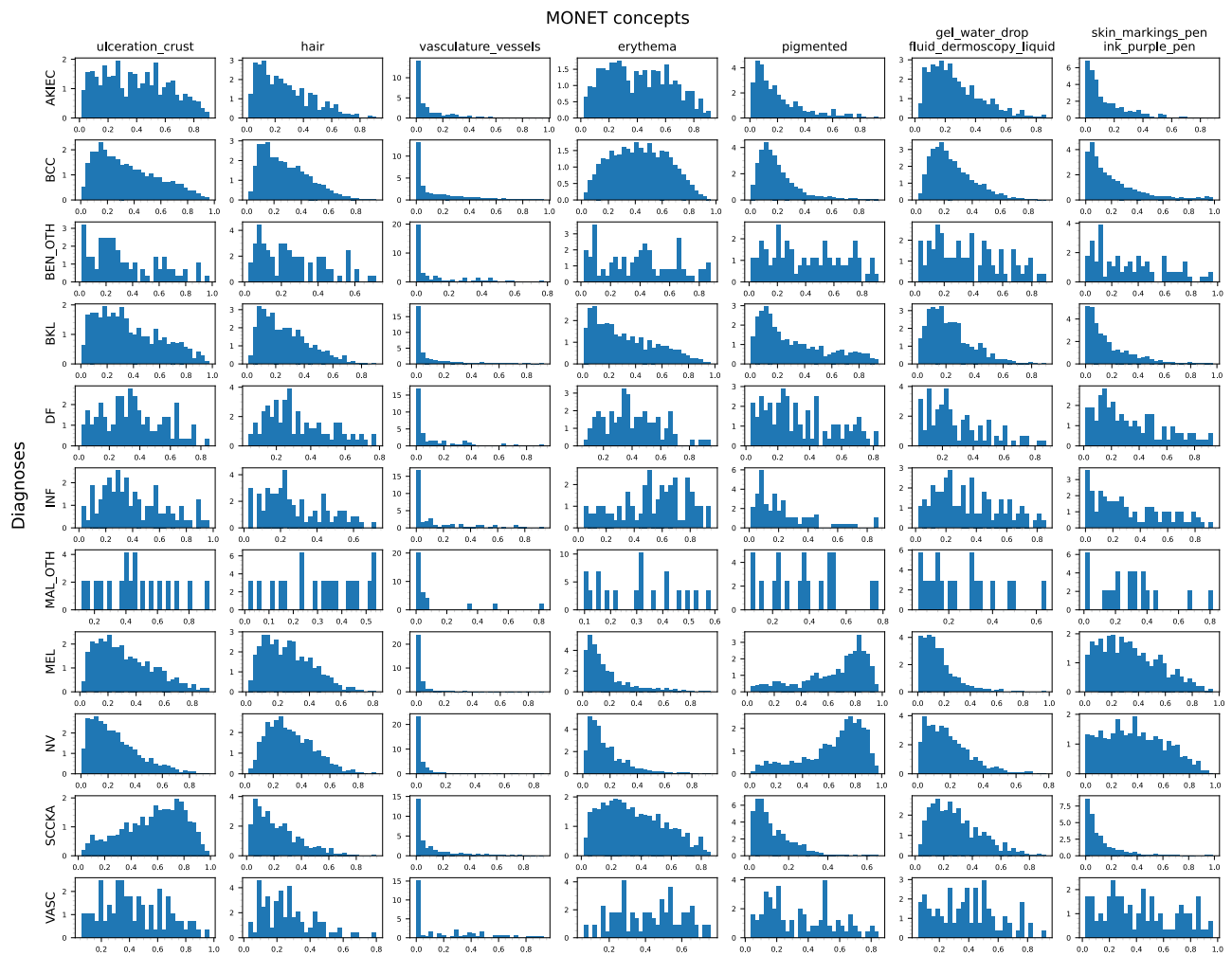
Supplementary information

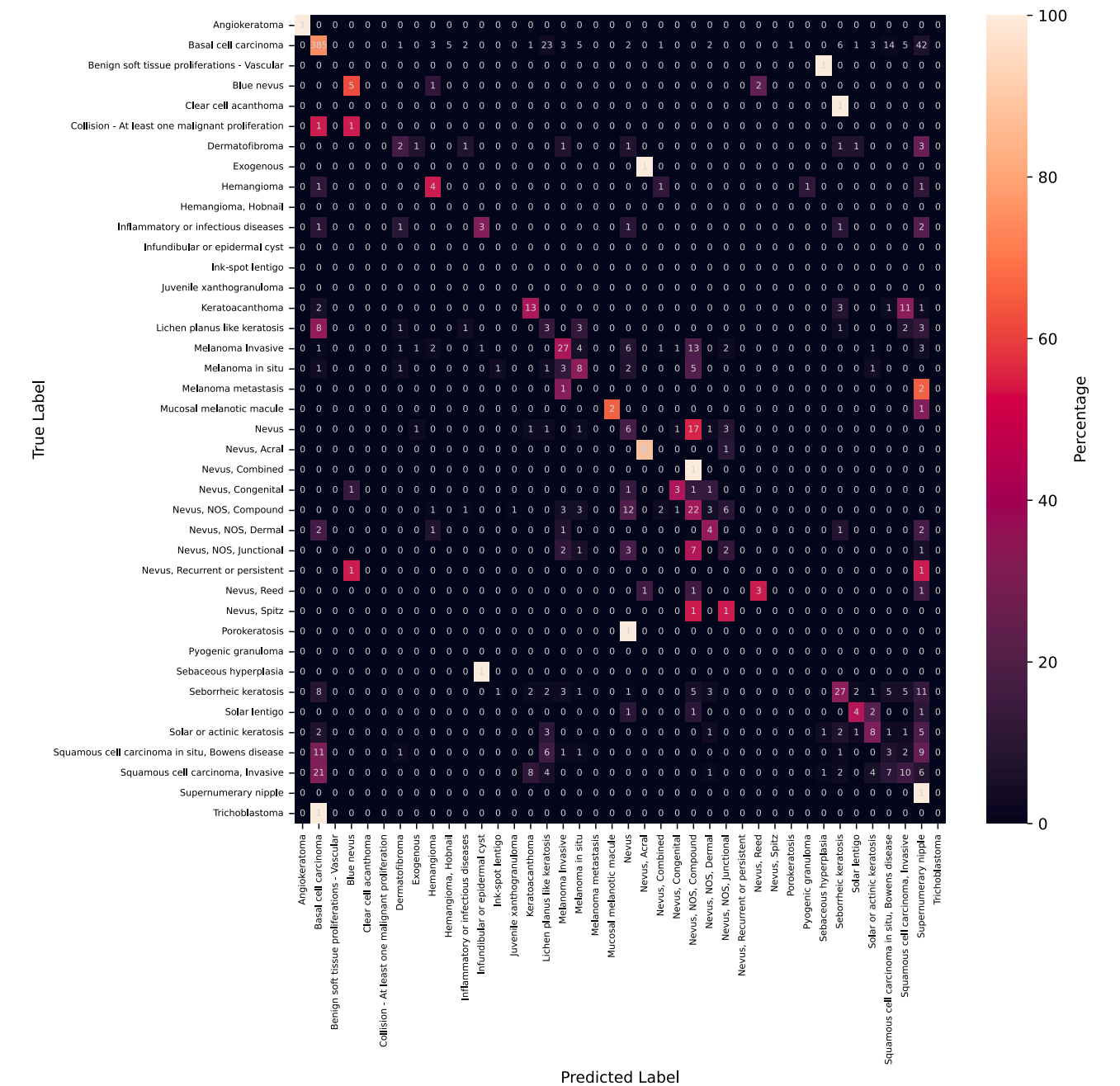
The machine-learning baseline code split the MILK10k dataset in 5 randomized cross-validation folds, stratified by the simplified ISIC2025 classes: a Siamese neural network, using ResNet50 as a backbone, a simple concatenation of the 2 penultimate layers, with a 512-dimensional projection layer, and a further fully connected layer toward the number of target classes, subject to 50% dropout during training. The target classes were the granular International Skin

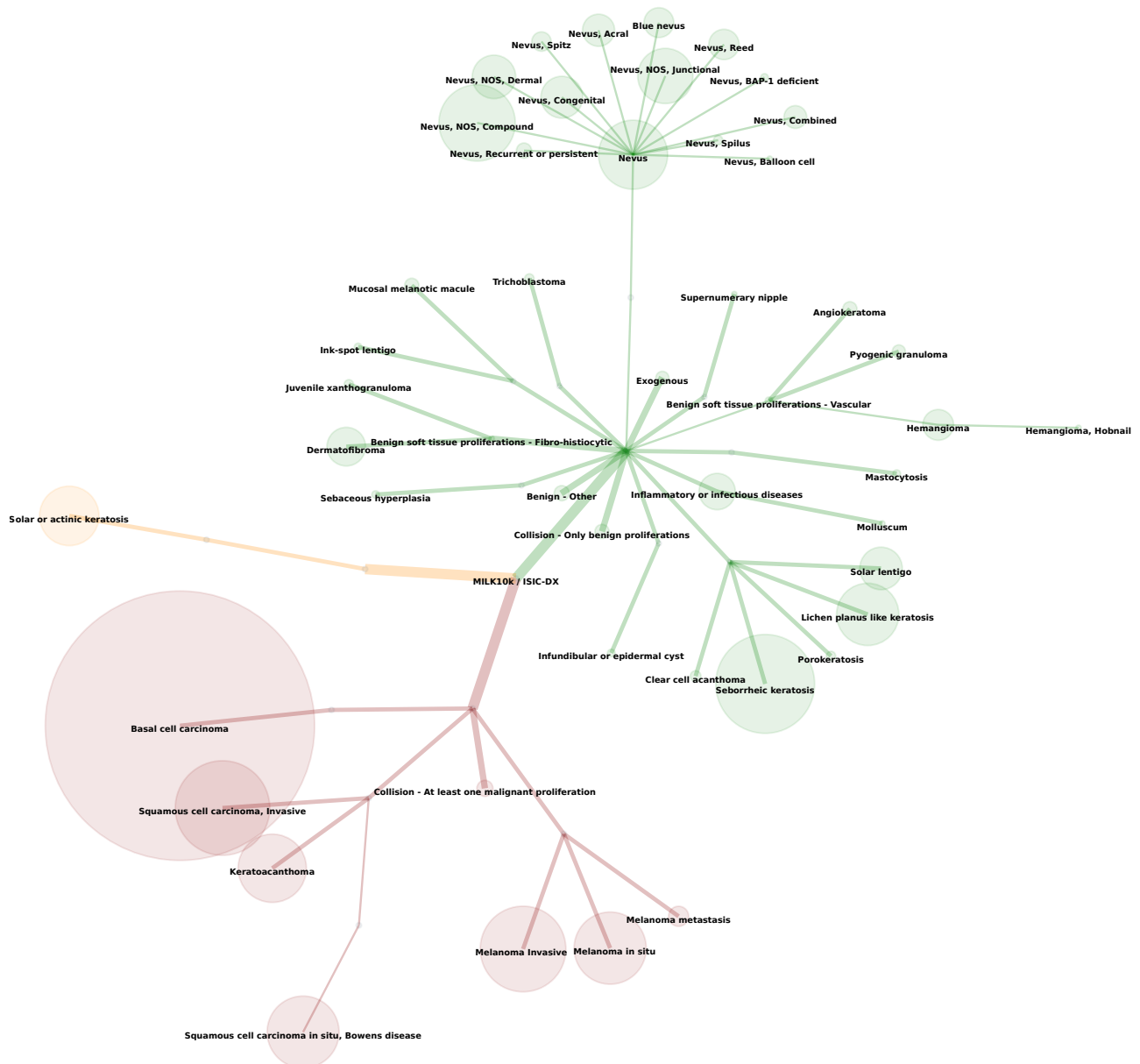
Imaging Collaboration-Designated Diagnoses classes, and finetuning was performed using CrossEntropyLoss, with label smoothing of 0.01 and a weighting of the loss based on the inverse class frequencies in the training split and AdamW as the optimizer. The learning rate was set at 0.0001, and the batch size was at 32, with step down of the learning rate in case of 4 epochs without validation loss improvement and early stopping after 10 epochs without improvement. The theoretical maximum of epochs was set as 100. Implementation details are provided through the link in the data availability statement.



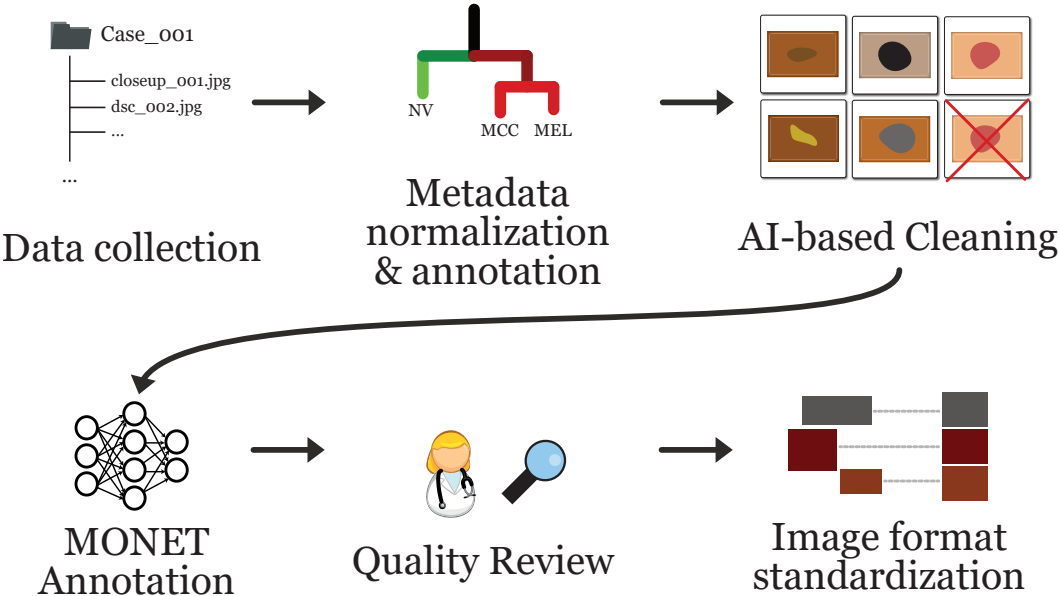
Supplementary Figure S1. ISIC-DX class coverage of ISIC 2017–2024 challenge raw datasets, available from the public ISIC-archive gallery API, DERM12345, and MILK10k. ISIC-DX, International Skin Imaging Collaboration-Designated Diagnoses.







Supplementary Figure S4. Graph representation of ISIC-DX–included classes in the MILK10k dataset. Colos represent benign (green), indeterminate (yellow), and malignant (red) classes; node sizes the number of included cases; and edge weight the distance from the root. ISIC-DX classes not present in MILK10k but part of paths between the root and diagnoses are visualized without text. ISIC-DX, International Skin Imaging Collaboration-Designated Diagnoses.



Supplementary Figure S5. Schematic workflow of processing the dataset.

Supplementary Table S1. Lesion Count Mapped to Granular ISIC-DX Cases and Corresponding Depth of the ISIC-DX Class within the Hierarchy

Diagnostic Class	ISIC-DX	Hierarchy Depth	n
AK/IEC	Solar or actinic keratosis	3	123
	Squamous cell carcinoma in situ, Bowen's disease	4	180
BCC	Basal cell carcinoma	3	2522
BEN_OTH	Benign - other	2	8
	Benign soft tissue proliferations - fibrohistiocytic	2	1
	Benign soft tissue proliferations - vascular	2	2
	Collision - only benign proliferations	2	7
	Exogenous	2	6
	Infundibular or epidermal cyst	3	2
	Juvenile xanthogranuloma	3	3
	Mastocytosis	3	2
	Mucosal melanotic macule	3	7
	Sebaceous hyperplasia	3	2
	Supernumerary nipple	3	1
	Trichoblastoma	3	3
BKL	Clear cell acanthoma	3	4
	Ink-spot lentigo	3	2
	Lichen planus-like keratosis	3	135
	Seborrheic keratosis	3	340
	Solar lentigo	3	63
DF	Dermatofibroma	3	52
INF	Inflammatory or infectious diseases	2	46
	Molluscum	3	1
	Porokeratosis	3	3
MAL_OTH	Collision - at least 1 malignant proliferation	2	9
MEL	Melanoma Invasive	3	255
	Melanoma in situ	3	181
	Melanoma metastasis	3	14
NV	Blue nevus	4	31
	Nevus	3	166
	Nevus, acral	4	36
	Nevus, BAP-1 deficient	4	2
	Nevus, balloon cell	4	1
	Nevus, combined	4	18
	Nevus, congenital	4	61
	Nevus, NOS, compound	4	203
	Nevus, NOS, dermal	4	66
	Nevus, NOS, junctional	4	106
	Nevus, recurrent or persistent	4	8
	Nevus, Reed	4	32
	Nevus, Spilus	4	3
	Nevus, Spitz	4	13
SCC/KA	Keratoacanthoma	3	162
	Squamous cell carcinoma, invasive	3	311
VASC	Angiokeratoma	3	7
	Hemangioma	3	33
	Hemangioma, hobnail	4	1
	Pyogenic granuloma	3	6

Abbreviations: AK, actinic keratosis; BKL, benign keratinocytic lesion; DF, dermatofibroma; IEC, intraepidermal carcinoma; ISIC-DX, International Skin Imaging Collaboration-Designated Diagnoses; KA, keratoacanthoma; SCC, squamous cell carcinoma.