

A General-Purpose Multimodal Foundation Model for Dermatology

Siyuan Yan^{1,2}, Zhen Yu¹, Clare Primiero³, Cristina Vico-Alonso⁴, Zhonghua Wang¹, Litao Yang^{1,2}, Philipp Tschandl⁵, Ming Hu^{1,2}, Gin Tan⁹, Vincent Tang¹⁰, Aik Beng Ng¹⁰, David Powell⁹, Paul Bonnington¹¹, Simon See¹⁰, Monika Janda⁶, Victoria Mar^{4,8}, Harald Kittler⁵, H. Peter Soyer^{*3,7}, Zongyuan Ge^{*1,2}

¹AIM for Health Lab, Faculty of Information Technology, Monash University, Melbourne, Australia

²Faculty of Engineering, Monash University, Melbourne, Australia

³Frazer Institute, The University of Queensland, Dermatology Research Centre, Brisbane, Australia.

⁴Victorian Melanoma Service, Alfred Hospital, Melbourne, Australia

⁵Department of Dermatology, Medical University of Vienna, Vienna, Austria.

⁶Centre for Health Services Research, Faculty of Medicine, The University of Queensland, Brisbane, Australia

⁷Dermatology Department, Princess Alexandra Hospital, Brisbane, Australia.

⁸School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia.

⁹eResearch Centre, Monash University, Melbourne, Australia

¹⁰NVIDIA AI Technology Center, Singapore

¹¹The University of Queensland, Brisbane, Australia

***Corresponding author:** Zongyuan Ge (zongyuan.ge@monash.edu) Peter Soyer (p.soyer@uq.edu.au)

Diagnosing and treating skin diseases require advanced visual skills across multiple domains and the ability to synthesize information from various imaging modalities. Current deep learning models, while effective at specific tasks such as diagnosing skin cancer from dermoscopic images, fall short in addressing the complex, multimodal demands of clinical practice. Here, we introduce PanDerm, a multimodal dermatology foundation model pretrained through self-supervised learning on a dataset of over 2 million real-world images of skin diseases, sourced from 11 clinical institutions across 4 imaging modalities. We evaluated PanDerm on 28 diverse datasets covering a range of clinical tasks, including skin cancer screening, phenotype assessment and risk stratification, diagnosis of neoplastic and inflammatory skin diseases, skin lesion segmentation, change monitoring, and metastasis prediction and prognosis. PanDerm achieved state-of-the-art performance across all evaluated tasks, often outperforming existing models even when using only 5-10% of labeled data. PanDerm’s clinical utility was demonstrated through reader studies in real-world clinical settings across multiple imaging modalities. It outperformed clinicians by 10.2% in early-stage melanoma detection accuracy and enhanced clinicians’ multi-class skin cancer diagnostic accuracy by 11% in a collaborative human-AI setting. Additionally, PanDerm demonstrated robust performance across diverse demographic factors, including different body locations, age groups, genders, and skin tones. The strong results in benchmark evaluations and real-world clinical scenarios suggest that PanDerm could enhance the management of skin diseases and serve as a model for developing multimodal foundation models in other medical specialties, potentially accelerating the integration of AI support in healthcare.

Introduction

There is a pressing need to fully harness the potential of artificial intelligence (AI) in diagnosing and managing skin diseases. Although deep learning has demonstrated remarkable performance, often matching or surpassing dermatologists, current AI models for dermatology remain limited to isolated tasks, such as diagnosing skin cancer from dermoscopic images¹. These models struggle to integrate diverse data types and imaging modalities, reducing their utility in real-world clinical settings. Dermatology, like internal medicine, is inherently complex, demanding a comprehensive, patient-centered approach. Diagnosing and treating skin cancer, for example, involves a range of tasks, including total body skin examination, risk assessment at both patient and lesion levels²⁻⁵, differentiation of neoplastic from inflammatory diseases⁶, multimodal image analysis^{7,8}, pathology interpretation^{9,10}, monitoring lesion changes^{11,12}, and predicting outcomes^{13,14}. The absence of integrated AI solutions capable of supporting these diverse workflows currently hampers the practical impact of AI in dermatology. Foundation models, however, hold the potential to address this gap by enabling a more holistic approach^{15,16}.

Foundation models are large-scale neural networks pretrained on vast, diverse datasets using self-supervised learning techniques, often leveraging weakly labeled or unlabeled data¹⁷⁻¹⁹. Built on rich knowledge representations, these models have demonstrated impressive performance across medical fields such as ophthalmology²⁰, radiology²¹, and pathology²²⁻²⁵. By leveraging large and diverse training datasets, these models are highly versatile in interpreting various data modalities²⁶, outperforming previous deep learning models in downstream tasks. Their strong feature representations also enable data-efficient applications^{27,28}, requiring fewer labeled samples, which is a crucial advantage in domains where labeled data are scarce.

The performance of foundation models is inherently linked to the scale of their parameters and training data²⁹⁻³¹. In general computer vision, foundation models are pretrained on massive datasets like ImageNet³² or JFT-300M³³ and most existing dermatology AI models still rely on these models for downstream adaptation. Some efforts have focused on self-supervised learning specifically for dermatology using public datasets^{34,35} or web-sourced skin images³⁶. However, these approaches are often limited by dataset size, diversity, or the lack of real patient data, which hinders the ability of models to generalize effectively across the diverse tasks and modalities encountered in clinical dermatology.

Here, we introduce PanDerm, a general-purpose, multimodal dermatology foundation model pretrained on over 2 million images sourced from 11 institutions across multiple countries, covering 4 imaging modalities (**Fig 1a-c**). This dataset represents the largest and most diverse image collection in dermatology to date. In the pretraining stage, PanDerm employs a novel combination of masked latent modeling and CLIP³⁷ feature alignment for self-supervised learning (**Fig 1e and Methods**), demonstrating superior data scalability and

training efficiency compared to existing self-supervised algorithms (**Fig 1f**). The model exhibits advanced capabilities in holistic patient analysis by effectively interpreting a wide range of dermatological imaging modalities, including total body photography (TBP) as well as clinical, dermoscopic, and dermatopathology images. This versatility supports patient analysis throughout clinical workflows focused on, but not limited to, skin cancer detection (**Fig 1d**).

We systematically evaluate PanDerm across 28 curated datasets (**Fig 1g**), covering a diverse array of clinical tasks, including screening, risk stratification, phenotype assessment, naevus counting, longitudinal monitoring, lesion change detection, diagnosis of neoplastic and inflammatory skin diseases, skin lesion segmentation, as well as recurrence prediction and prognosis. PanDerm achieves state-of-the-art performance on all tasks, often using only 5-10% of the labeled training data typically required. Furthermore, we demonstrate PanDerm’s clinical impact through three real-world studies, including malignant lesion detection using total body photography, human-AI collaboration, and early melanoma detection. These results collectively underscore PanDerm’s potential to enhance dermatological practice and may extend to other medical specialties, potentially accelerating the development and clinical integration of AI across various healthcare domains.

Results

Ablation studies and comparison with other self-supervised learning strategies

To evaluate PanDerm’s effectiveness, we conducted ablation studies to examine the impact of pretraining data size and training epochs on downstream performance (datasets described in **Extended Data Table 28**). PanDerm demonstrated strong scalability, with consistent AUROC improvements as pretraining data increased from 0.8 to 1.8 million images (**Fig 1f left**). Notably, PanDerm outperformed SwAVDerm³⁶, a recent dermatology self-supervised learning model, using only 0.8 million images compared to SwAVDerm’s 1.2 million (**Fig 1f left**). Furthermore, PanDerm demonstrated superior training efficiency compared to state-of-the-art self-supervised learning models, achieving better performance with only 200 training epochs, while MILAN³⁹ required 500 epochs, and both DINOv2⁴⁰ and MAE¹⁹ needed 800 epochs (**Fig 1f right**), all using the same pretraining dataset. This efficiency stems from PanDerm’s novel use of CLIP as a teacher model for semantic feature learning. PanDerm also surpassed vision-language models such as CLIP³⁷, MONET⁴¹, and BiomedCLIP⁴² in benchmark evaluations (**Extended Data Table 28**). Additionally, It demonstrated emergent capabilities in dermatology similar to those of DINOv2 in natural images, with linear probing performance comparable to full-parameter fine-tuning (**Extended Data Table 29**). Building on these promising initial results, we extended our evaluation to a broader range of dermatological tasks, primarily comparing PanDerm with three representative pretrained models in the following sections: SL-Imagenet³² and DINOv2⁴⁰ (both widely-used foundation models pretrained on natural images with a ViT-large³⁸ backbone), and SwAVDerm³⁶

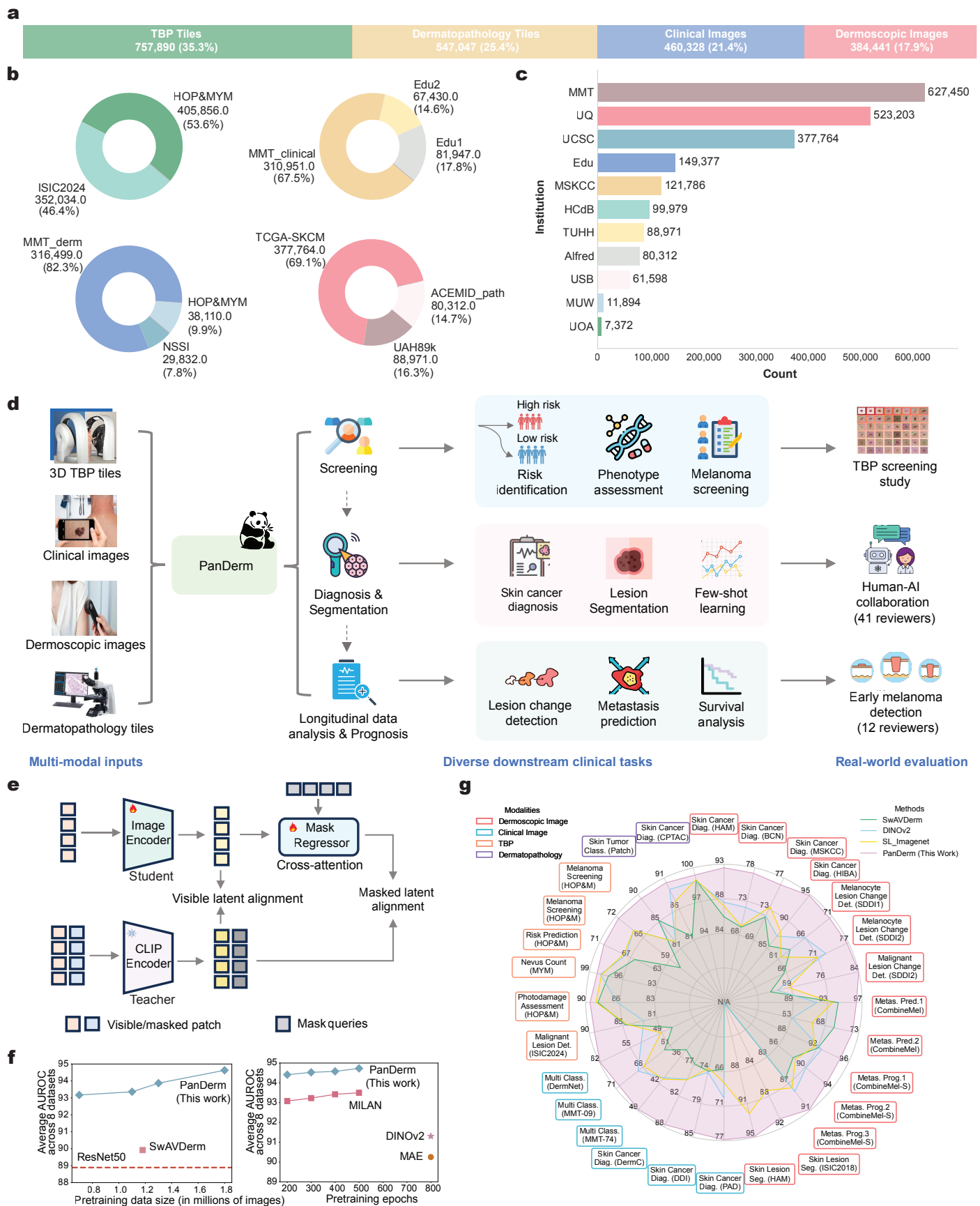


Figure 1: Overview of this study. Caption on next page.

(Previous page.) **Figure 1: Overview of this study. a-c.** The pretraining dataset includes 2.1 million images from 11 clinical sources and institutions across 4 modalities, representing the largest and most diverse multi-modal dermatology dataset to date. Dataset distribution is shown by modalities (a), data sources (b), and institutions (c). **d.** PanDerm can interpret various imaging modalities and perform a wide range of dermatology-related clinical tasks. It has been evaluated in a real-world scenario study (e.g., TBP-based melanoma screening) and two reader studies (human-AI collaboration and early melanoma detection). Dermatopathology images: microscopic images of skin biopsy specimens at various magnifications. Clinical images: wide-field or regional images capturing the lesion and surrounding skin taken with digital cameras. Dermoscopic images: close-up images taken with a dermoscope, showing detailed lesion structures. TBP tiles: lesion crops taken from macro TBP images. **e.** PanDerm features a decoupled architecture comprising a ViT-large³⁸ encoder, a regressor, and a CLIP-based teacher model. Its pretraining process employs two objectives: a latent representation reconstruction objective and a CLIP latent alignment objective. **f.** PanDerm’s performance across varying pretraining data sizes and epochs, measured by average AUROC on 8 benchmark datasets, with comparative analysis against other pretraining strategies. **g.** PanDerm surpasses existing pretrained models on 28 evaluation datasets, covering various clinical tasks across 4 modalities.

(a self-supervised model pretrained on a large skin image dataset from search engines).

Diagnostic performance and generalization ability over various datasets

We systematically evaluated PanDerm diagnostic performance across 10 public datasets from 4 imaging modalities and 7 international sites (**Fig 2a**). These datasets covered multi-class classification of pigmented neoplastic lesions and binary melanoma diagnosis tasks. We primarily evaluated performance using weighted F1 score for multi-class datasets and area under the receiver operating characteristic curve (AUROC) for binary class datasets. PanDerm consistently outperformed all other models, achieving significant improvements on 9 out of 10 datasets, with average gains of 5.1%, 8.0%, 4.2%, and 0.9% on dermoscopic, clinical, TBP, and pathology datasets, respectively (**Fig 2a**). On benchmarks like HAM10000³⁵ and PAD⁴³, PanDerm surpassed the next-best models by 4.7% ($P < 0.001$) and 9.0% ($P < 0.001$), respectively (**Fig 2a; Extended Data Table 1 and Extended Data Fig 3**).

When evaluating label efficiency generalization, PanDerm consistently outperformed other models across all datasets (**Fig 2b, Extended Data Table 15-20**). PanDerm matched the next-best model’s performance using only 10% to 30% of labeled data for skin lesion diagnosis across various modality datasets (**Fig 2b**), demonstrating PanDerm’s significant advantage in limited data scenario. Additional results for other tasks are presented in **Extended Data Fig 4**. To further assess PanDerm’s generalization ability, we extended a previous out-of-distribution generalization experiment^{44,45} to evaluate model performance for melanoma diagnosis on both dermoscopic and close-up clinical images from 7 external datasets. These datasets were collected from multinational institutions, representing populations distinct from the training data. PanDerm demonstrated significant superiority over all pretrained models, achieving higher AUROC scores across all external datasets (**Fig 2c**). Notably, PanDerm showed a significant improvement even on clinical images not used during fine-

tuning, with AUROC gains of 4.0%, 2.6%, and 2.1% on the three clinical datasets (All $P < 0.001$).

We further evaluated PanDerm’s ability to classify a broader range of skin conditions, encompassing up to 74 classes. These included not only pigmented skin lesions but also inflammatory, infectious, and rare diseases. The evaluation was conducted on three clinical image datasets: the public DermNet dataset ⁴⁶ (23 classes), and two in-house datasets, MMT-09 (9 classes) and MMT-74 (74 classes). As shown in **Fig 2d**, PanDerm achieved weighted F1 improvements of 3.2%, 7.1%, and 8.2% on MMT-09, DermNet, and MMT-74, respectively, compared to the next-best models (all $P < 0.001$). We observed PanDerm’s performance advantage increased as the number of skin conditions grew, indicating a strong generalization and discriminative capacity across diverse diseases. PanDerm also outperformed all other pretrained models on all metrics across the three datasets (all $P < 0.001$; **Extended Data Table 2**). In the DermNet dataset, PanDerm exceeded the next-best model’s area under the precision-recall curve (AUPR) by 14.7%. Further details on the experimental setup, datasets, and metrics are provided in **Methods**.

Reader study 1: Early melanoma detection compared with clinicians

We conducted a reader study to compare PanDerm’s performance in early melanoma detection using sequential images. We used a dermoscopic image dataset from Alfred Hospital ⁴⁷, featuring multiple follow-up images of the same lesions over time. The study evaluated two key aspects: overall diagnostic accuracy and early melanoma detection capability. PanDerm’s performance was compared to that of 12 human reviewers (7 experienced dermatologists and 5 registrars). In terms of overall accuracy, PanDerm outperformed the average human reviewer by 10.2% and surpassed the best-performing human by 3.6%. For early detection, we assessed the earliest time point of correct malignant diagnosis for each melanoma case. PanDerm demonstrated superior ability in this challenging task, correctly identifying 77.5% (69 out of 89) of melanoma lesions at the first imaging time point, compared to only 32.6% (29 correct diagnoses) for human reviewers. Individual dots in the histograms of **Fig 2e** represent the earliest correct diagnosis time points for both PanDerm and human reviewers, visualizing the comparative early detection performance. Further details on the reader study setup and datasets are provided in **Methods**.

Lesion monitoring and change detection in sequential images

Monitoring suspicious melanocytic lesions over a three-month period is a widely accepted procedure for early melanoma detection, as changes often prompt excision to rule out melanoma, whilst stability can be reassuring ¹². We assessed PanDerm’s ability for short-term lesion change detection by framing it as a binary classification task: determining whether a pair of short-term lesion images shows change or remains stable. A Siamese network architecture ⁴⁸, which is well-suited for comparing image pairs to detect subtle changes, was employed.

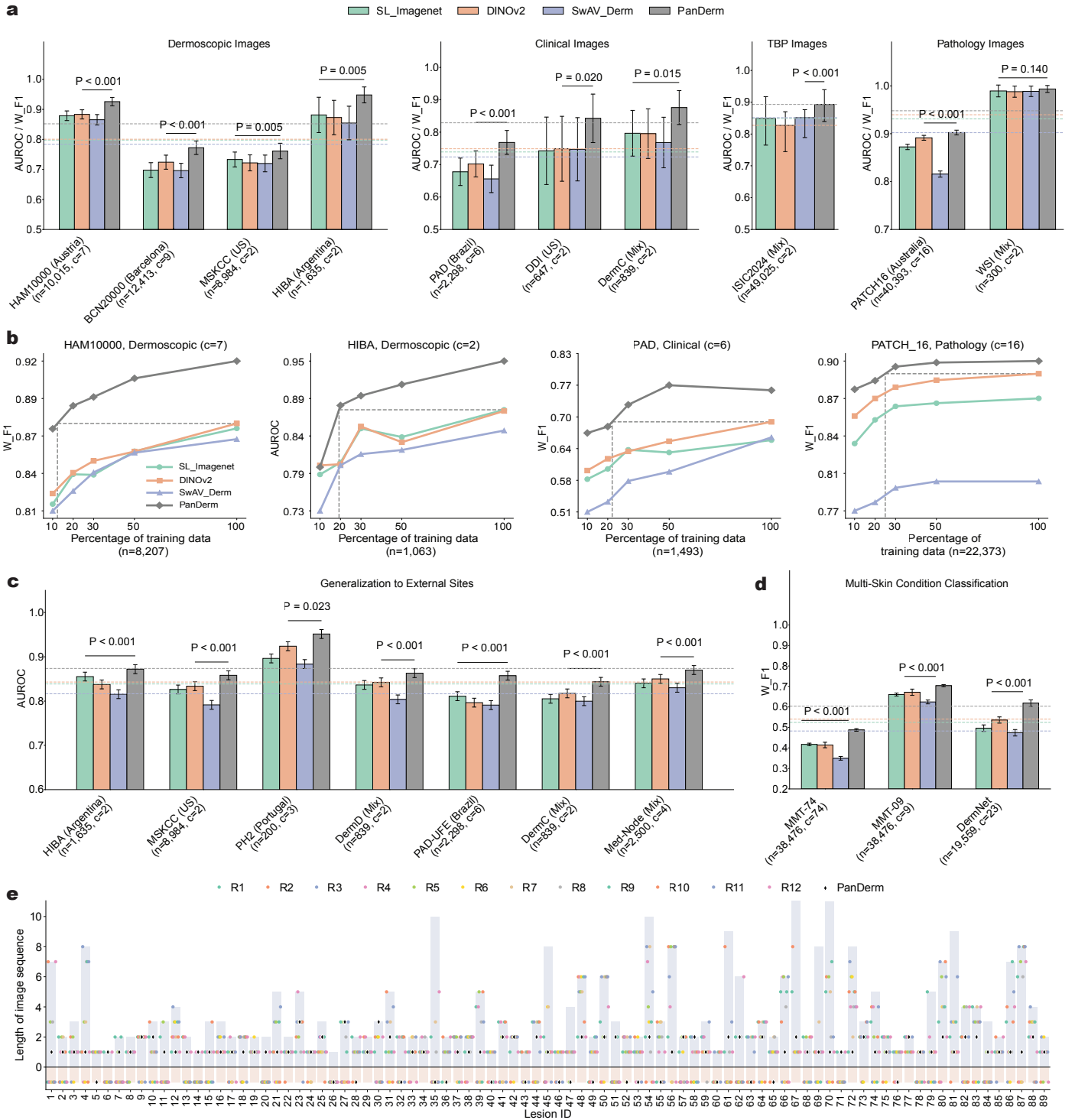


Figure 2: PanDerm's versatile capacity in diverse diagnosis tasks. Caption on next page.

(Previous page.) **Figure 2. PanDerm’s versatile capacity in diverse diagnosis tasks.** **a.** Performance of PanDerm versus other pretrained models on 10 pigmented skin lesion datasets across multiple centers and modalities. *n*: data size, *c*: class number. Metrics: AUROC for binary class datasets ($c=2$) and W_F1 (weighted F1) score for multi-class datasets ($c>2$). Dashed lines indicate the average performance of each model across different datasets. **b.** Comparison between PanDerm and other pretrained models in a label efficiency generalization setting on four representative datasets, demonstrating model performance at various percentages of training data. Vertical dash lines show the data quantity needed for PanDerm to match existing model performance. **c.** External validation for melanoma diagnosis across 7 datasets. **d.** Performance evaluation of general skin condition classification (up to 74 classes) using clinical images. **e.** Early melanoma detection results (reader study 1) comparing PanDerm to 12 clinicians (7 experienced dermatologists, 5 registrars). X-axis: 89 melanoma lesion IDs; Y-axis: lesion image sequence length. Points on the histogram represent the initial time points of correct melanoma diagnoses. Points below $y=0$ correspond to melanoma lesions undetected throughout the sequence. Error bars in **a**, **c**, **d** show 95% CIs; bar centers in **a**, **c**, **d** represent mean value; dots in **b** represent mean value. Estimates are computed using nonparametric bootstrapping with 1000 bootstrap replicates. *P*-values are calculated using a two-sided t-test.

The primary challenges in this task are the subtlety of changes and the variability in imaging conditions. To address these challenges, we developed a comprehensive image pre-processing and alignment pipeline that includes dark corner removal, skin inpainting, image registration, and segmentation (**Fig 3c**). We then compared PanDerm’s performance on two longitudinal dermoscopic datasets, SDDI1 and SDDI2 (**Fig 3a**, **Fig 3b**, and **Methods**). Ablation studies (**Fig 3d left**) showed that our pre-processing pipeline significantly improved PanDerm’s performance on lesion monitoring, increasing AUROC from 0.596 (95% CI 0.567-0.624) to 0.706 (95% CI 0.686-0.725) in SDDI1 ($P < 0.001$) and from 0.683 (95% CI 0.517-0.894) to 0.767 (95% CI 0.649-0.886) in SDDI2 ($P < 0.001$). Using the optimized pipeline for all models, PanDerm achieved AUROC improvements of 4.3% in SDDI1 ($P < 0.001$) and 3.7% in SDDI2 over the next-best model (**Fig 3d, middle**). For lesions later diagnosed as malignant (**Fig 3b, middle**), PanDerm achieved an AUROC of 0.840 (95% CI 0.769-0.911), surpassing the next-best model by 15.0% ($P < 0.01$) (**Fig 3d, right**). Further details on lesion change detection are provided in **Methods** and **Extended Data Table 3-5**.

Metastasis prediction and prognosis

We evaluated PanDerm’s potential to identify digital biomarkers for melanoma progression using dermoscopic images, a relatively underexplored but clinically valuable area^{13,14,49} (**Fig 3f**). The evaluation used the Combin-Mel dataset, an international multi-center cohort with 680 dermoscopic images of invasive primary melanoma from 370 patients. PanDerm’s performance was assessed for predicting melanoma metastasis (**Fig 3g**, details in **Methods**). In binary classification (control vs. metastasis), PanDerm achieved an AUROC of 0.964 (95% CI 0.937-0.991), outperforming the next-best model by 2.0% ($P = 0.073$) (**Fig 3e**). For the more challenging three-class classification (control vs. local metastasis vs. distant metastasis), PanDerm significantly outperformed the next-best model by 2.8% ($P < 0.05$) in weighted F1 score (**Extended Data Table 6**).

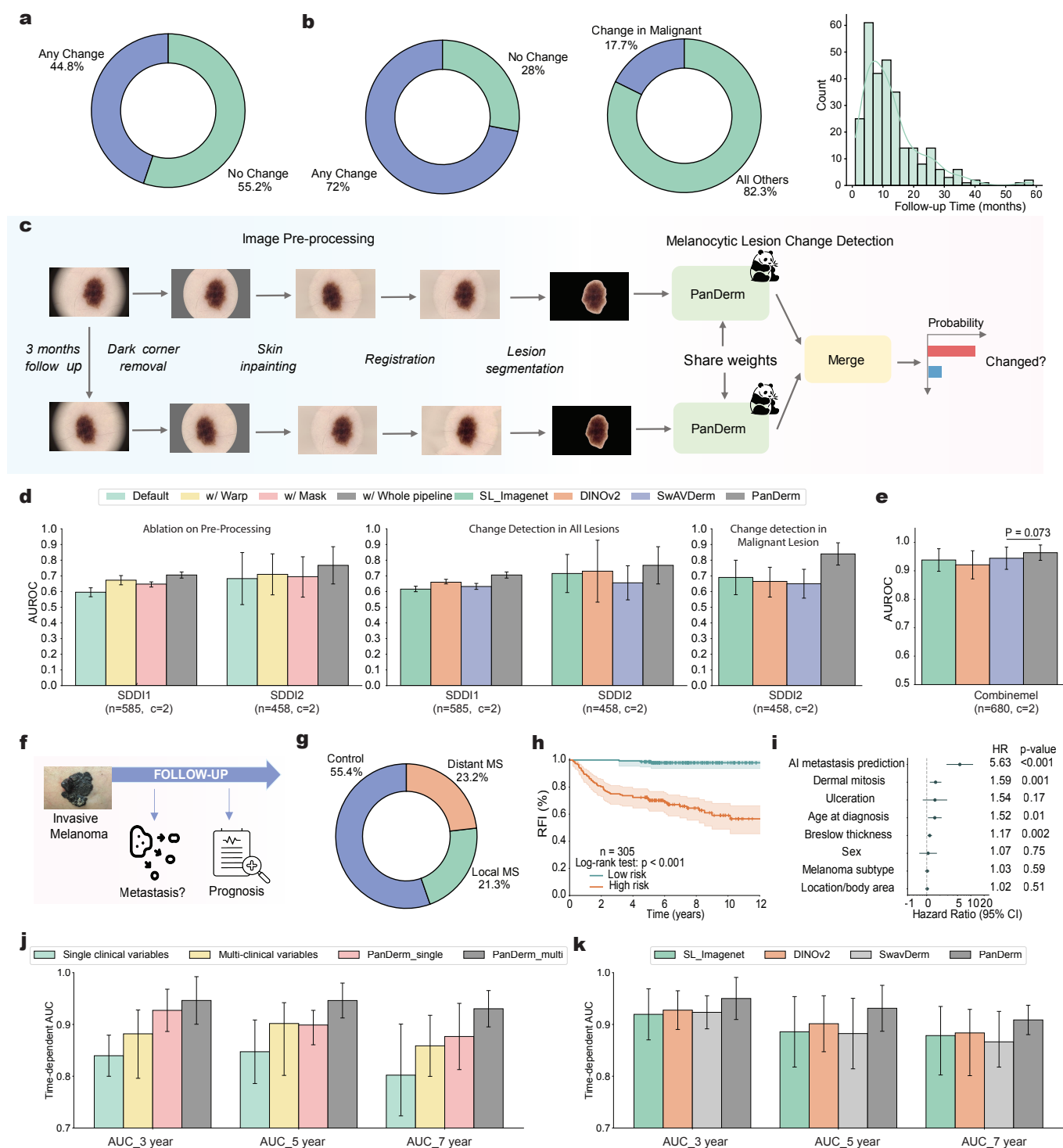


Figure 3: Short-term lesion change detection, metastasis prediction, and prognosis results. Caption on next page.

(Previous page.) **Short-term lesion change detection (a-d) and metastasis prognosis (e-k) results.** **a.** SDDI1 dataset statistics for melanocytic lesion change detection: ratio of changed lesions during follow-up, ratio of changed malignant lesions during follow-up, and follow-up time distribution. **b.** Ratio of changed lesions in the SDDI2 dataset. **c.** Longitudinal dermoscopic image-based lesion change detection using PanDerm. For comparing subtle changes in paired lesions during short-term follow-up (e.g., 3 months), images undergo dark corner detection and removal, skin inpainting, registration, and lesion segmentation. This allows models to focus on subtle differences between lesions at different time points. **d.** Ablation study on pre-processing: “Default” (direct input of sequential images), “w/Warp” (registration only), “w/Mask” (lesion segmentation only), and “w/Whole pipeline”(complete pre-processing as in **c**). For change detection in all lesions and malignant lesions, all models are evaluated using the same whole pre-processing pipeline for fair comparison. **e.** Performance of binary metastasis prediction (control vs. metastasis), measured by AUROC. **f.** Scheme of PanDerm for predicting melanoma metastasis and prognosis. **g.** Distribution of metastasis types in the Combinemel dataset (MS represents metastasis). **h.** Kaplan–Meier curves for the recurrence-free interval (RFI) in invasive melanoma patients (CombinMel dataset), stratified by PanDerm prediction scores. **i.** Forest plots of hazard ratios for PanDerm, stratified groups in invasive melanoma patients. **j.** Time-dependent AUC of PanDerm vs. clinical variable score combinations. **k.** Time-dependent AUC comparison of PanDerm and other pretrained models. Error bars in **d-e** and **j-k** represent 95% CIs; bar centers indicate the mean value. Estimates are computed with five-fold cross-validation.

To evaluate the prognostic value of PanDerm’s metastasis prediction scores, we conducted survival analyses using Kaplan-Meier analysis and Cox proportional hazards regression. Kaplan-Meier analysis (**Fig 3h**) demonstrated that patients stratified into the high-risk group based on PanDerm’s prediction scores had significantly shorter recurrence-free intervals (RFI) compared to those in the low-risk group (HR: 5.63, 95% CI: 2.87-11.02, $P < 0.001$). We then performed multivariate Cox regression, incorporating PanDerm’s prediction scores alongside clinical variables such as sex, age, Breslow thickness, ulceration, dermal mitosis, location, and melanoma subtype. The resulting hazard ratios (**Fig 3i**) indicated that PanDerm’s metastasis prediction was the strongest predictor of RFI among all variables considered. To further assess predictive accuracy, we evaluated the time-dependent AUC at different intervals, comparing PanDerm’s score with single and multiple clinical variables (**Fig 3j**). PanDerm achieved AUCs of 0.950 (95% CI 0.910-0.991), 0.931 (95% CI 0.887-0.976), and 0.909 (95% CI 0.880-0.937) at 3, 5, and 7 years, outperforming multi-clinical variables by 6.8%, 2.9%, and 5.0%, respectively. Combining PanDerm’s score with clinical variables further improved AUCs at 5 and 7 years (**Fig 3j**). PanDerm also outperformed all pretrained models (**Fig 3k**), exceeding the next-best model (DINOv2) by 2.3%, 3.0%, and 2.5% at 3, 5, and 7 years, respectively. Details on the experimental setup and datasets are provided in **Methods**.

Skin phenotype, risk assessment and malignant lesion screening using TBP

We evaluated PanDerm’s performance in skin phenotype, risk assessment, and malignant lesion screening using 3D TBP^{2,3,50} (**Fig 4a**). Unlike dermoscopy, TBP allows for the integration of broader patient-level data, facilitating the assessment of risk factors such as photodamage and nevus count, both critical for melanoma

risk assessment ^{4,5,51}. Using in-house TBP data, PanDerm achieved a weighted F1 score of 0.896 (95% CI 0.879-0.913) for photodamage assessment and an AUROC of 0.983 (95% CI 0.979-0.987) for nevus counting, outperforming all other models ($P < 0.05$ and $P < 0.001$, respectively; **Fig 4b, c, g**). Notably, PanDerm maintained its performance advantage even with limited labeled data, outperforming the next-best model while using only 10% of the training data (**Extended Data Fig 4**). In lesion-specific risk stratification, PanDerm also ranked first with an AUROC of 0.705 (95% CI 0.698-0.712) and BACC of 0.657 (95% CI 0.6513-0.663), with all results statistically significant ($P < 0.001$; **Fig 4d, h**).

For skin cancer screening using TBP, PanDerm effectively identified malignant lesions among a large number of benign lesions in a highly imbalanced dataset (216 malignant vs. 197,716 benign lesions) from the HOP ⁵² and MYM ⁵³ cohort (**Fig 4e**). PanDerm outperformed the next-best model in sensitivity by 4.2% when using only TBP data (**Fig 4j left**), and by 3.5% when integrating additional measurement data (**Fig 4j right**), achieving a sensitivity of 0.893. It successfully detected malignant lesions in 79 out of 80 patients and recommended significantly fewer benign lesions for dermoscopy compared to melanographers (3498 vs. 8913), thereby reducing unnecessary examinations by approximately 60.8% (**Fig 4j, k; Extended Data Table 10**).

The distribution of lesions, visualized through UMAP plots (**Fig 4f**), demonstrated that PanDerm’s feature space effectively differentiates suspicious lesions, aligning with the clinical “ugly duckling” concept ⁵⁴. This concept involves identifying atypical lesions by comparison with other lesions on the same individual. Clustering patterns (i.e., clusters of lesions with similar characteristics) in PanDerm’s risk stratification (**Fig 4l**) corresponded closely with those observed in human screenings (**Fig 4i**), illustrating PanDerm’s exceptional performance in malignant lesion screening. Additional details on the experimental setup, dataset, measurement information and variable importance for TBP, and performance metrics are available in the **Methods** section, **Extended Data Table 7-10**, and **Extended Data Fig 5**.

Skin lesion segmentation

We evaluated PanDerm’s performance on skin lesion segmentation using two benchmark datasets: ISIC2018 (task 1) ⁵⁵ and HAM10000 ³⁵. PanDerm was compared against SL-Imagenet, autoSMIM ³⁴, and BATFormer ³⁴. PanDerm significantly outperformed all models, surpassing the next-best by 3.1% and 1.9% in Jaccard index (JAC) on ISIC2018 and HAM10000, respectively ($P < 0.001$; **Extended Data Fig 1a, b**). Its Dice score (DSC) was also significantly higher than that of all other models across both datasets ($P < 0.001$). PanDerm’s performance was particularly noteworthy in label-limited scenarios, matching the next-best model while using only 5% of the training data (104 and 350 images for ISIC2018 and HAM10000, respectively; **Extended Data Fig 1c, d**). To further validate its capabilities, PanDerm was compared with MedSAM ⁵⁶, a state-of-the-art medical segmentation foundation model, and outperformed it in JAC by 0.5% ($P = 0.025$ and 0.112;

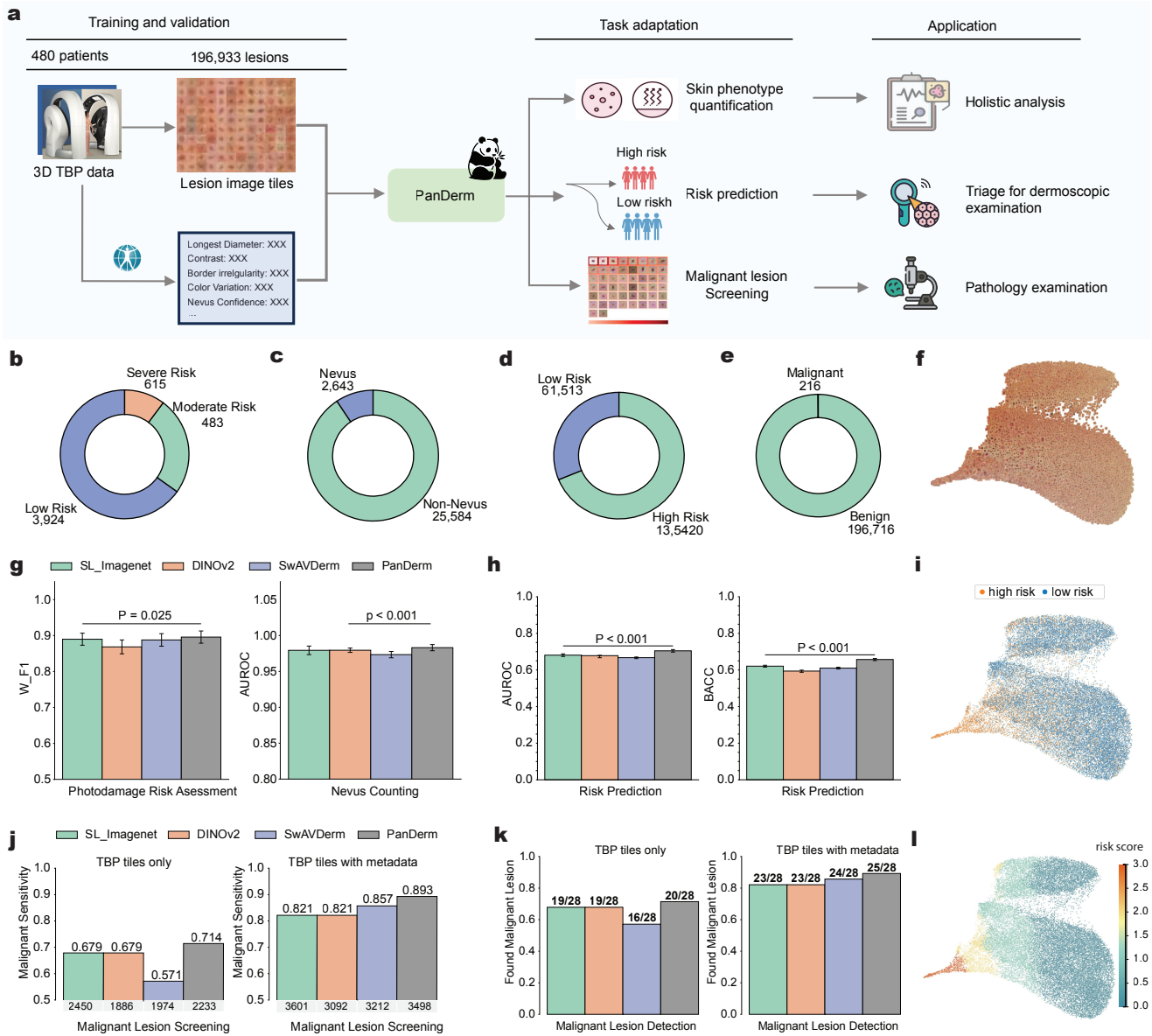


Figure 4: Skin phenotype assessment and melanoma screening using TBP. **a.** Illustration of PanDerm processing multimodal TBP data (HOP & MYM dataset) for total body skin examination, enhancing skin phenotype score quantification, risk stratification, and malignant lesion screening for advanced early melanoma detection. **b-c.** Class distribution of skin phenotype for solar damage risk ((b)) and nevus count ((c)) in the HOP and MYM datasets. **d-e.** Class distribution of risk groups and malignant lesions. **g.** Solar damage risk assessment and nevus counting performance, measured by W_F1 (weighted F1) and AUROC, respectively. **h.** Risk prediction performance, measured by AUROC and BACC (Balanced Accuracy). Error bars in **g-h** show 95% CIs; bar centers represent mean value. The estimates are computed with nonparametric bootstrapping using 1000 bootstrap replicates. P -values are calculated with a two-sided t-test. **j.** Malignant lesion screening performance, measured by sensitivity for malignant lesions. Left: using only Total Body Photography (TBP) data. Right: integrating measurement information. Numbers below each bar indicate the recommended count of suspicious lesions. **k.** Number of malignant lesions detected in the test set. **f.** UMAP plot of PanDerm screening results for all test lesions. **i.** UMAP plot of human screening results for all test lesions. **l.** UMAP plot of PanDerm risk prediction results for all test lesions.

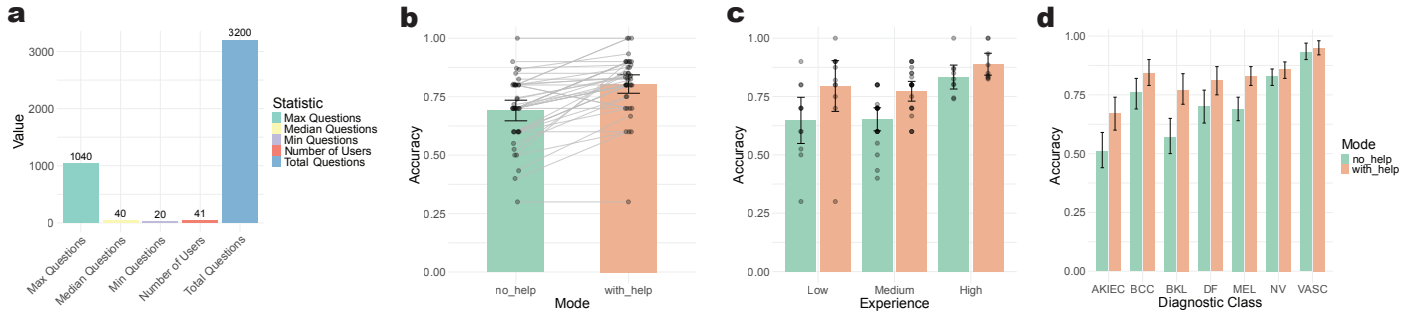


Figure 5: Performance of PanDerm in a Human-AI collaboration setting. **a.** Overview of the reader study: 41 users participated, answering a total of 3,200 questions. The maximum, median, and minimum number of questions answered per user were 1,040, 40, and 20, respectively. **b.** Comparison of diagnostic accuracy without support and with PanDerm support ($P < 0.001$; two-sided paired t-test; $n = 41$ readers). **c.** Comparison of diagnostic accuracy without support and with PanDerm support, grouped by experience level: Low ($n = 11$ readers), Medium ($n = 21$ readers), and High ($n = 9$ readers). **d.** Comparison of diagnostic accuracy without support and with PanDerm support, grouped by diagnostic classes. Abbreviations: MEL, melanoma; BCC, basal cell carcinoma; AKIEC, actinic keratosis/intraepidermal carcinoma; BKL, benign keratinocytic lesion; NV, melanocytic nevus; DF, dermatofibroma; VASC, vascular lesion.

Extended Data Table 13) on the two datasets. This is notable since PanDerm does not use segmentation-specific datasets during pretraining, unlike MedSAM. Additionally, PanDerm uses a smaller input size (224×224) compared to MedSAM (1024×1024), resulting in significantly lower computational costs. On an RTX3090 GPU, PanDerm’s fine-tuning and inference speeds were approximately 5 and 4 times faster than MedSAM’s, respectively (**Extended Data Table 14**). Further details on segmentation tasks are provided in **Methods**, with qualitative results in **Extended Data Fig 2** and quantitative results in **Table 11-14**.

Reader study 2: Human-AI collaboration

To assess PanDerm’s clinical applicability, we conducted a reader study to investigate its potential to improve clinicians’ diagnostic accuracy. The study included 41 human raters with diverse experience levels. We compared their accuracy in diagnosing seven classes of pigmented lesion classes using dermoscopic images, both with and without PanDerm’s multi-probability prediction support (**Fig 5a**). PanDerm’s support significantly increased overall diagnostic accuracy from 0.69 (95% CI 0.65-0.73) to 0.80 (95% CI 0.76-0.84, $P < 0.001$; **Fig 5b**). A subgroup analysis revealed that raters with less experience benefited the most, showing accuracy improvements of 17% ($P = 0.0082$) for those with low experience and 12% ($P < 0.001$) for those with medium experience, while highly experienced raters showed a 6% improvement ($P = 0.039$; **Fig 5c**, **Extended Data Table 25**). The class-specific analysis demonstrated significant accuracy improvements in 4 out of the 7 classes ($P < 0.05$; **Fig 5d**, **Extended Data Table 24**). For melanoma specifically, PanDerm improved the accuracy of human raters from 0.69 (95% CI 0.64-0.74) to 0.83 (95% CI 0.79-0.87, $P < 0.001$). Additional details on the reader study can be found in the **Methods** section.

Discussion

Despite significant advances in AI technology, its application in clinical medicine remains fragmented and underutilized. Current AI systems are often restricted to isolated tasks, unable to address the diverse demands of medical decision-making. This limits AI’s potential in supporting clinicians in disease diagnosis and management. Dermatology, with its complex requirements including holistic patient assessment, lesion-specific analysis, and potential use of various imaging modalities, serves as an ideal use case for demonstrating AI’s capabilities across multiple interconnected clinical tasks. Success in this domain could pave the way for broader adoption of AI models across healthcare.

In this study, we introduce PanDerm, a versatile dermatology foundation model trained through self-supervised learning on over 2 million multimodal dermatological images. Central to PanDerm’s development was the curation of a large and diverse image dataset sourced primarily from in-house collections and carefully selected public repositories. This approach contrasts with previous efforts, such as SwAVDerm³⁶, which relied on web-sourced skin data, inadvertently incorporating images from commonly used benchmarks like ISIC⁵⁷ and DermNet⁴⁶, increasing the risk of data leakage and compromising evaluation validity. Our strategy minimizes this risk, ensuring that benchmark evaluations accurately reflect real-world model performance.

To assess PanDerm’s potential for extensive clinical applications, we evaluated the foundation model across 28 benchmark datasets encompassing the most common dermatology-related clinical tasks. PanDerm outperformed existing models in skin cancer-related tasks, including risk stratification of individual phenotypes and lesions, detection of lesion changes and malignancy, and metastasis prediction and prognosis. Moreover, it demonstrated superior performance in diagnosing neoplastic and inflammatory skin diseases. Notably, PanDerm achieved these results often using significantly less training data (5-10%) than existing models, addressing the critical challenge of acquiring high-quality labeled data that typically require specialist input. Reader studies further validated PanDerm’s real-world applicability, providing evidence of improved diagnostic accuracy with AI support. Collectively, these findings underscore PanDerm’s transformative potential, enhancing AI integration into clinical workflows.

The scaling behavior observed in PanDerm’s performance, particularly in relation to pretraining dataset size, aligns with recent trends in foundation model development^{20,22,23}. However, our study uniquely demonstrates this phenomenon in dermatological AI, where large-scale data acquisition poses distinct challenges. The positive impact of increasing dataset size and diversity on model performance underscores the importance of collaborative data collection efforts across institutions and countries. Our comparative analysis also provides key insights into the effectiveness of different self-supervised learning approaches. Using CLIP³⁷ as a teacher model for semantic feature alignment resulted in superior training data efficiency (**Fig 1f**), significantly outper-

forming DINOv2 ⁴⁰. This is especially relevant in healthcare, where assembling datasets comparable to those used in natural image processing is often impractical. For context, DINOv2 typically requires massive datasets like LVD-142M (142 million images), a scale that is rarely achievable in medical domains. Furthermore, our masked feature reconstruction approach demonstrated a stronger capacity for capturing semantic-level information compared to methods like MAE ¹⁹, which is a critical capability in the diagnosis of skin diseases, where subtle semantic features can be diagnostically significant. These advantages allowed PanDerm to consistently outperform traditional models such as ResNet50 ⁵⁸ and ViT large ³⁸, as well as recent generalist medical foundation models such as Med-PaLM-M ⁵⁹ and BiomedGPT ⁶⁰ across a broad spectrum of clinical tasks (**Extended Data Table 30**). The superior performance of PanDerm over generalist medical AI models can be attributed to several factors. Despite their broad scope, current generalist models struggle to achieve clinical-grade performance in specialized medical domains. This underperformance stems from limitations such as insufficient domain-specific data, lack of coverage for specialized modalities like TBP, and challenges in acquiring large, high-quality datasets across diverse medical domains. Additionally, potential modality conflicts ⁶¹ and negative transfer issues ⁶⁰ between distinct medical domain data further hinder their performance in highly specialized clinical tasks. In contrast, PanDerm’s focused approach to skin diseases overcomes these limitations, resulting in more clinically relevant performance.

Furthermore, PanDerm demonstrated robust performance across diverse demographic factors, including different body locations, age groups, genders, and skin tones, setting it apart from previous models (**Extended Data Table 21-22**). Notably, minimal bias was observed across different skin tones, contrasting sharply with prior studies ⁶²⁻⁶⁴ that reported significant disparities in AI performance for dermatological conditions among diverse populations. This equitable performance, attributed to PanDerm’s diverse training data and semantic-focused approach, has important implications for reducing health disparities and promoting equitable healthcare delivery in dermatology. PanDerm’s demonstrated clinical applicability in real-world scenarios marks a significant step towards integrating AI support into clinical practice. The model showed considerable promise in enhancing early melanoma detection, facilitating melanoma screening, and improving diagnostic accuracy when used in collaboration with clinicians. These capabilities align closely with the vision of AI as an augmentative tool in healthcare ⁶⁵⁻⁶⁷.

Despite these promising results, our study has limitations that warrant further investigation. Currently, PanDerm focuses primarily on imaging data and tasks related to skin cancer diagnosis. Integrating large language models ^{29,68} and additional data modalities, such as medical reports, proteomics, genomics, and microarray data, into PanDerm’s training could further enhance its diagnostic capabilities and comprehensive understanding of complex medical information. Additionally, while our training data is diverse, its geographic concentration may not fully represent global variation. Future work should prioritize expanding the geographic and ethnic diversity of the training dataset to improve global applicability. In the future, developing foundation

models for skin disease diagnosis should emphasize expanding data collection through international institutional collaborations, such as those led by the International Skin Imaging Collaboration (ISIC) ⁵⁷, while ensuring careful consideration of data quality and potential biases in large-scale sources. Such efforts will be crucial for advancing dermatological AI and enhancing its clinical utility across diverse populations worldwide.

In conclusion, PanDerm represents a significant advancement providing a robust and versatile multi-modal foundation model for the diagnosis of skin diseases. Its development addresses key challenges in the field, including data curation, model architecture, training strategy, and clinical validation. By leveraging a large-scale, diverse dataset and advanced self-supervised learning techniques, PanDerm demonstrates superior performance across a broad range of dermatological tasks, consistently outperforming all comparative pretrained models. The principles and methodologies established in this study could serve as a roadmap for developing foundation models in other medical specialties, potentially accelerating the integration of AI support in healthcare.

Methods

Pretraining dataset for developing PanDerm

We curated an extensive pretraining dataset comprising 2,149,706 unlabeled multimodal skin images to develop PanDerm. This diverse dataset encompasses 4 imaging modalities and 11 data sources, including total body photography (TBP), dermatopathology, clinical, and dermoscopic images. The composition of the dataset is as follows: 757,890 (35.3%) TBP tiles, 537,047 (25.4%) dermatopathology tiles, 460,328 (21.4%) clinical images, and 384,441 (17.9%) dermoscopic images. The inclusion of multiple imaging modalities aims to provide a comprehensive representation of skin lesions and conditions, enabling the model to learn robust features across different visual representations. This large-scale dataset serves as the foundation for pretraining PanDerm, allowing it to capture the intricate patterns and characteristics of various skin conditions across different imaging techniques.

MYM & HOP cohort (TBP). The MYM cohort ⁵³ is an in-house dataset designed to study the natural history of melanocytic naevi. This cohort recruited individuals from the general population who were 18 years or older and had at least one mole. The HOP study ⁵² is an in-house sequential dataset targeting high-risk melanoma individuals. Inclusion criteria for the HOP study were: at least one melanoma (including in situ) diagnosed before the age of 40 years, two or more melanomas (including in situ) diagnosed before the age of 65, a strong family history (two or more first-degree relatives affected) and/or known pathogenic mutations in a hereditary melanoma gene, and/or a diagnosis of dysplastic naevus syndrome. For both cohorts, 3D TBP was conducted using a VECTRA WB360 (Canfield Scientific Inc., Parsippany, NJ, USA), which instantaneously captures 92

cross-polarised 2D images with standardized lighting and subsequently merges them to create a 3D avatar. The average number of lesion tiles per subject from TBP was approximately 500. Demographic factors were collected using standard questionnaires, with clinical characteristics collected by research assistants. The final dataset of automatically detected lesion image tiles ≥ 2 mm in diameter comprises 405,856 images. More detailed information about these two datasets can be found in our previous studies ^{52,53}.

MYM & HOP cohort (dermoscopic). The MYM and HOP datasets also contain 38,110 dermoscopic images from suspicious lesions identified during the studies. These images provide a complementary view of the lesions of interest, offering detailed visualization of surface and subsurface structures that may be indicative of various skin conditions, particularly melanoma.

MMT dataset. The MMT dataset is an in-house collection amassed from over 150 clinics across Australia and New Zealand over a 15-year period. This extensive dataset primarily consists of paired polarized dermoscopic and clinical images. From this comprehensive collection, we curated a subset containing 316,399 dermoscopic images and 310,951 clinical images, providing a rich source of pretraining data for training purposes.

ACEMID pathology pilot study. The ACEMID dermatopathology dataset is an in-house collection comprising 98 slides, which have been processed to yield 80,312 patch images.

NSSI. The NSSI dataset is an in-house sequential collection containing 29,832 dermoscopic images. These images were collected as part of the Brisbane Naevus Morphology Study, conducted from approximately 2009 to 2014. The images were captured using a digital dermatoscope attached to a Fotofinder ATBM imaging system, resulting in images of approximately 0.05MB, with dimensions of 768×576 pixels at 96 dpi. The dataset includes up to 7 time points per participant, scheduled every 6 months over a 3-year period, with some variation due to late enrollment, scheduling non-adherence, or loss to follow-up. Individual lesions were assigned a consistent number across visits, facilitating temporal analysis.

Edu1 & Edu2. The Edu1 and Edu2 datasets are curated from in-house educational resources, such as educational notes and materials. These datasets contain 81,947 and 67,430 clinical images, respectively, providing a diverse range of examples that are typically used for training medical professionals in dermatology.

ISIC2024. ISIC2024 ⁵⁰ is an open-source TBP-based dataset designed for identifying skin cancers among skin lesions cropped from 3D total body photographs. For our pretraining purposes, we selected a subset of the dataset, stratified by institutions, containing 352,034 tile images.

TCGA-SKCM. The TCGA-SKCM dataset ⁶⁹ is derived from The Cancer Genome Atlas (TCGA) project, which characterized the mutational landscape of human skin cutaneous melanoma (SKCM). This dataset con-

tains 475 slides, which have been processed into 377,764 patch images.

UAH89k. The UAH89k dataset ⁷⁰ is a subset of a larger collection, comprising 269 histopathology whole slide images (WSIs) sourced from the archives of the Institute of Pathology, Heidelberg University, the MVZ for Histology, Cytology and Molecular Diagnostics Trier, and the Institute for Dermatopathology. This dataset provides additional histopathological data, further enriching the model’s understanding of skin conditions at the microscopic level.

Detail of model architecture and pretraining

PanDerm is a self-supervised learning model designed for the dermatology field, built upon the success of existing self-supervised learning techniques in the natural image domain ⁷¹. At its core, the architecture comprises a ViT-Large visual encoder ³⁸, a mask regressor, and a CLIP-Large ³⁷ teacher model. The ViT-Large encoder, with its 24 transformer blocks and 1024 dimensional embeddings, processes 224×224 -pixel images, while the CLIP-Large teacher model handles slightly smaller 196×196 -pixel inputs. The training process incorporates two primary objectives: masked latent alignment and visible latent alignment loss. Initially, the input image undergoes masking, with the mask ratio proportional to the encoder’s complexity (50% for ViT-Large). The encoder then processes visible patches to produce latent representations, while the regressor predicts the latent representations of masked patches using these visible latent and mask tokens. The model focuses on the encoder-regressor structure without a separate decoder component. The regressor assumes the responsibility of predicting the latent representations of masked patches, allowing for more efficient processing and learning. For target supervision, the unmasked image is fed through the CLIP model, generating supervision divided according to visible and masked patch locations. The visible latent alignment loss is directly applied to the latent representations of visible patches computed by the encoder. Concurrently, the masked latent alignment loss acts on the latent representations of masked patches predicted by the regressor. Both of these loss functions use CLIP latent representations as their supervision signals. The regressor in PanDerm operates similarly to a cross-attention mechanism. It uses learnable mask tokens as queries, while the keys and values are derived from the concatenation of visible patch representations and the output of previous layers. This design allows the regressor to effectively infer the content of masked regions based on the context provided by visible areas. Optimization primarily focuses on aligning the visible and masked patch predictions with their corresponding CLIP latent supervisions. This approach enables PanDerm to extract rich, semantically meaningful representations from dermatological images without relying on explicit labels.

For pretraining, we continued to train the model (initially trained on ImageNet-1K) on our dataset of over 2 million unlabeled multimodal skin images, representing diverse dermatological conditions. We set the batch size on each GPU to 480, with an effective batch size of 1920. Following masked image modeling practices ⁷²,

we used a 50% mask ratio. To pretrain our model, we used AdamW as the optimizer with an initial learning rate of $1.5e-3$. We apply simple data augmentation such as random resized cropping and horizontal flipping during pretraining. We trained our model for 500 epochs with a warmup of 20 epochs. The pretraining phase used 4 80GB NVIDIA H100 GPUs and took approximately 5 days and 7 hours. We chose the last epoch checkpoint as our final model weights. Please refer to **Extended Data Table 31** for more detailed pretraining hyperparameter configurations.

Target representations (Teacher model) of PanDerm. Prior work^{39,71,72} demonstrated that target representations produced by the teacher model for masked image modeling are essential for impacting model performance. We ablated different teacher models, including two widely used models that demonstrated promising performance (CLIP-base and CLIP-large), a biomedical domain-specific CLIP (BiomedCLIP⁴²), and a dermatology-specific CLIP (MONET⁴¹). We observed that CLIP-large pretrained on the natural domain can outperform biomedical-specific CLIP and dermatology-specific CLIP. This can be attributed to the limited data scale of skin images in medical domain CLIP models. Thus, CLIP-large remains the best teacher model for creating target representations for masked image modeling in dermatology. Based on these findings, we selected CLIP-large as the teacher model; the performance of our model when incorporating CLIP-large teachers was significantly improved and also outperformed CLIP-large itself. Please refer to **Extended Data Table 28** for detailed results.

Linear probing vs fine-tuning for PanDerm. One emergent capacity of foundation models in the natural image domain is that the model’s features are ready for downstream tasks without needing to fine-tune the encoder model, such as in DINOv2⁴⁰. We explored whether PanDerm could achieve this capacity despite having different training objectives. We found that our model using simple linear probing can perform comparably with expensive full-parameter fine-tuning. This suggests that PanDerm’s features are already well-suited for diverse downstream multimodal skin-related tasks without requiring further training. Detailed results are in **Extended Data Table 29**.

Downstream Evaluation Details

Competing self-supervised learning baselines. For self-supervised learning methods comparison, we primarily evaluated DINOv2⁴⁰, MAE¹⁹, and MILAN³⁹, all utilizing the same ViT-large backbone. We employed the recommended hyperparameter configurations for these models and continued pretraining from their natural image training weights on our pretraining dataset. Subsequently, we fine-tuned these models using identical hyperparameter setups to ensure a fair comparison.

Fine-tuning and linear probing. In adapting PanDerm to downstream tasks, only the encoder model is

utilized. For most tasks, PanDerm’s feature quality suffices to achieve competitive performance using simple linear probing. This involves applying a linear classifier (i.e., logistic regression) to the top of extracted features from the PanDerm encoder to evaluate its performance on downstream tasks. For more challenging tasks requiring higher performance, we opted to fine-tune the PanDerm encoder. The fine-tuning tasks include the two reader studies, short-term change detection, skin lesion segmentation, skin cancer detection in ISIC2024, and TBP-based risk stratification. For all other tasks, we employed linear probing. For linear probing, following practices recommended by the self-supervised learning community, we fix the ℓ_2 regularization coefficient λ to $MC/100$, where M is the embedding dimension and C is the number of classes, and employ the L-BFGS solver with a maximum of 1,000 iterations. For fine-tuning, we adhere to the BEiT V2 setting⁷², utilizing cross-entropy loss with a learning rate of 5×10^{-4} . We train models for 50 epochs with a warmup of 10 epochs. The model exhibiting the best performance on the validation set is selected as the final model. For detailed hyperparameter configurations, please refer to **Extended Data Table 32**. In the following sections, we describe tasks with more specific methodological details.

Sequential data preprocessing for lesion change detection. Our proposed sequential data preprocessing method consists of dark corner removal, skin inpainting, hair removal, image registration, and lesion segmentation. For the first two steps, we follow the approach outlined in ⁷³. Given an image with or without dark corner artifacts (DCA), we first convert it to grayscale. We then extract the contour by applying OpenCV’s ⁷⁴ binary threshold function to the grayscale image, empirically setting the threshold at 100, and using the findContours function with RETR_TREE mode and CHAIN_APPROX_SIMPLE method. We identify the largest area contour in the image, which most closely matches the edge of the DCA, by calculating the area of all existing contours. Using OpenCV’s minEnclosingCircle function, we capture a circular area that encompasses this largest contour. To mitigate the effect of gradient colors at the dark corner edges, we scale this circle down to 80% of its original radius and convert it into a binary mask. Finally, we inpaint the original image using this mask, employing OpenCV’s implementation of the Telea algorithm with an inpaint radius of 10. Following dark corner removal and inpainting, we implement a hair removal step to further improve image quality and facilitate more accurate registration. This process begins by converting the image to grayscale. We then apply a black hat morphological operation using a 17×17 structuring element to isolate dark, thin structures (hairs) from the background. The resulting image is thresholded to create a binary mask of the detected hair structures. Finally, we use OpenCV’s inpaint function with the Telea algorithm to fill in the hair regions, effectively removing them from the image. This hair removal step is crucial for improving the accuracy of subsequent image registration and analysis. For image registration, we implement an AKAZE ⁷⁵ feature-based approach. The process begins by detecting key points and computing descriptors using the AKAZE algorithm, which is particularly effective for non-linear scale-spaces. We utilize OpenCV’s AKAZE_create function, setting the descriptor size to 0, threshold to 9×10^{-5} , and number of octaves to 4. Keypoint matching is performed using a Brute Force matcher with Hamming distance and cross-checking enabled. To refine the matches and esti-

mate the transformation between images, we employ the RANSAC (Random Sample Consensus) algorithm implemented via `skimage.measure.ransac`. This estimates an `EuclideanTransform` model, which accounts for rotation and translation between the images. The resulting transformation is then applied to warp one image onto the other using `skimage.transform.warp` with reflection padding and linear interpolation.

Siamese network for change detection. Similar to ⁴⁸, we employ a simple Siamese network architecture for change detection, where two identical visual encoders with shared weights from our foundation model process a pair of sequential lesion images captured over a short time frame. Each encoder extracts features from its respective image. These learned features are then concatenated and passed through two fully connected layers, followed by a softmax layer for final classification. For training this siamese network in our binary change detection task, we use a contrastive loss function. This loss is particularly well-suited for Siamese networks as it helps the model learn to distinguish between pairs of images that have changed and those that have not. The contrastive loss encourages the network to minimize the distance between feature representations of image pairs with no significant changes while maximizing the distance for pairs that show meaningful changes. This approach allows the network to learn a similarity metric between image pairs, rather than simply classifying individual images. By doing so, it becomes more sensitive to subtle changes between images and more robust in detecting clinically relevant lesion changes over time. The contrastive loss thus helps the model focus on learning features that are most relevant for distinguishing between changed and unchanged lesion pairs, improving its overall performance in change detection tasks.

Early melanoma detection (Reader study 1). We fine-tuned our foundation model on the private SDDI-Alfred dataset⁴⁷ using a ten-fold cross-validation approach. We utilized cross-entropy loss with a learning rate of 5×10^{-4} . We train models for 50 epochs with a warmup of 10 epochs. The model exhibiting the best AUROC on the validation set is selected as the final model. We then employed an out-of-fold (OOF) prediction approach to generate melanoma predictions for all sequential images. For each image sequence, we recorded the time point at which the model first made a correct diagnosis of melanoma; otherwise, the model was considered to have failed in detecting the melanoma. For the human evaluation, 12 clinicians—seven dermatologists with over five years of experience and five registrars with less than five years of experience—were invited to assess the serial dermoscopic data. The images were presented to the reviewers using Qualtrics™ (Provo, UT, USA), with the reviewers blinded to the true diagnoses. For each case, information such as the patient’s age, sex, lesion location, and date of imaging was provided. Initially, only the first dermoscopic image in the sequence was shown, and reviewers were asked to classify the lesion as either benign or malignant. As they progressed through the sequence, side-by-side image comparisons were made available to assess changes over time. Once a diagnosis was submitted, it could not be revised. To mitigate bias, we included 10 single time-point melanoma images, preventing reviewers from assuming that the first image in a series was benign. We then compared the diagnostic performance of the clinicians with our model, focusing on the time point at which a malignant

diagnosis was first made by either the clinicians or the algorithm.

Melanoma metastasis prediction and survival analysis. We employ a linear probing classifier on our foundation model to predict melanoma metastasis using dermoscopic images from the private CombinMel dataset. Our evaluation encompasses two scenarios: binary metastasis prediction and multi-class metastasis prediction. In the binary classification, we aim to differentiate between the presence of any metastasis (including local/satellite/in-transit metastasis, lymph node recurrence, and distant metastasis) and its absence. The multi-class prediction presents a more complex challenge, categorizing cases into three groups: control (no metastasis), local/satellite/in-transit metastasis, and distant metastasis. To enhance the robustness and mitigate potential data selection bias, we perform five iterations of dataset splitting into training and testing sets, stratified by melanoma stage. The model is trained using this five-fold data. We linear probe PanDerm with the setting mentioned above. We then generate out-of-fold (OOF) predictions for all lesions and compare these to the ground truth for performance evaluation.

Subsequently, we conduct a multivariate Cox regression analysis, incorporating the metastasis prediction score and clinical variables (age, sex, Breslow thickness, ulceration, dermal mitosis, melanoma subtype, and lesion location) to predict the recurrence-free interval (RFI). This analysis focuses on earlier stages of melanoma (stages I-II). We visualize the relative contribution of individual variables to prognosis prediction using a forest plot. To analyze the correlation between variables and RFI, we employ the Kaplan-Meier method. Patients are stratified into low-risk and high-risk groups based on their binary metastasis prediction scores (median value). The log-rank test is utilized to assess the classifier’s ability to predict survival. To evaluate the predictive accuracy at various time points, we generate time-dependent receiver operating characteristic (ROC) curves and calculate areas under the curve (AUCs) at 3, 5, and 7 years. This approach allows us to assess the model’s performance over different follow-up periods.

Melanoma screening using TBP. The melanoma screening algorithm is designed to identify high-risk lesions among whole-body images, aiding clinicians in efficiently detecting potential malignancies. Lesions flagged as high-risk undergo further triage and dermoscopic examination. The screening model integrates three modules: a risk prediction head, an ugly duckling (UD) detection head, and a machine learning (ML) module, utilizing both TBP image data (image tiles) and metadata for comprehensive predictions. We first fine-tune our foundation model, equipped with the risk prediction head, using TBP image tiles to classify lesions as high-risk or low-risk. All lesion images are resized to 224×224 pixels and subjected to data augmentation, including color and geometric transformations. The risk prediction head, comprising a single linear layer, identifies lesions as high-risk if subjected to dermoscopy examination and low-risk otherwise. The UD detection head leverages the “Ugly Duckling sign”, an effective diagnostic strategy that compares all lesions from the same patient to identify outliers. This approach capitalizes on lesion contextual information. We use the fine-tuned foundation

model to extract deep learning features, which are then processed by the UD detection head. This module calculates the distance between each lesion’s features and the average features of all lesions from the same patient, employing the interquartile range (IQR) method to select outlier lesions. The ML module, an extra tree classifier, is trained using TBP metadata, which includes 32 measurements for each lesion from the 3D TPB machine. This module directly predicts malignancy based on pathology labels. The final screening result combines predictions from all three modules. A lesion is flagged as suspicious for malignancy if any module yields a positive prediction. We evaluate the screening performance at both the lesion and patient levels to ensure comprehensive accuracy assessment.

Weakly supervised slide classification. Weakly supervised slide classification tasks are approached using the established two-stage multiple instance learning (MIL) framework. This process encompasses: 1) extraction of instance-level features from discrete tissue regions within the whole slide image (WSI), and 2) development of an adaptable, order-invariant aggregation method to consolidate patch-level data into a comprehensive slide-level representation. For slide preprocessing, we employ the CLAM toolbox⁷⁶, utilizing its built-in parameters for tissue segmentation. The resulting regions are partitioned into 256×256 non-overlapping sections at 20 \times magnification or its equivalent. During the feature extraction phase, these sections undergo resizing to 224×224 and normalization using ImageNet mean and standard deviation parameters. To ensure consistency, all pretrained encoders utilize an identical set of patch coordinates for feature extraction across each WSI.

To evaluate the efficacy of various pretrained encoders in weakly-supervised learning, we implement the Attention-Based Multiple Instance Learning (ABMIL) algorithm⁷⁷ with consistent architectural configurations across all comparisons. Our implementation features a two-tier gated ABMIL structure, where the initial fully connected (FC) layer maps input embeddings to a 512-dimensional space, followed by intermediate layers with 384 hidden units. To enhance generalization, we incorporate dropout regularization, applying rates of 0.10 and 0.25 to the input embeddings and subsequent intermediate layers, respectively. The training protocol employs the AdamW optimizer⁷⁸ with a cosine learning rate schedule, initializing the learning rate at $1e-4$ and setting weight decay to $1e-5$. We utilize cross-entropy as our loss metric. The training process is limited to 20 epochs, implementing an early stopping mechanism based on validation loss performance. To ensure robust evaluation, we employ a five-fold cross-validation strategy, stratifying our slide dataset by both case and label attributes.

Skin lesion segmentation. For skin lesion segmentation, we employ a conventional segmentation paradigm, utilizing a network encoder connected to a segmentation decoder and head. Our proposed PanDerm serves as the encoder in this setup. We benchmark PanDerm against three established models: ViT-large³⁸, autoSMIM³⁴, and BATFormer⁷⁹. Both ViT and PanDerm use a UperNet decoder, following the official ViT implementation. For autoSMIM and BATFormer, we adhere to their official repository settings. ViT-large and autoSMIM encoders are initialized with ImageNet pretrained weights. To ensure a fair comparison, all images are resized

to 224×224 . We apply online data augmentation, including color jittering, random rotation, and random flipping, to mitigate overfitting. The training employs an AdamW optimizer with an initial learning rate of $5e-4$ and a weight decay of 0.01, with the learning rate decaying according to a cosine schedule. The models are trained for 100 epochs, and we save the model that achieves the best evaluation metrics on the validation set.

Interaction platform, raters, reader study and implementation for human-AI collaboration. The reader study was conducted using DermaChallenge, a web-based platform developed and hosted by the Medical University of Vienna for online education on dermatoscopy, as described in a previous study⁶⁵. To ensure proper authentication and data management, readers were required to register with a unique username, valid email address, and password. Active users on the platform, who previously actively agreed to be contacted, were recruited via a single email. Before commencing to the study phase, all users had to finish three introduction levels to be familiarized with the platforms’ user interface and image types. The number of correct answers in the first iteration of these levels normalized against the mean score of the entire DermaChallenge platform user base, served as a score of experience. Users were grouped into “Low” ($n=11$), “Medium” ($n=21$), and “High” ($n=9$) experience based on quantiles with cuts at 0.25 and 0.75 probability (R *stats::quantile()* function). Within the study level, users were shown batches of 10 images, randomly selected from a pool of 1,511 images, i.e. the ISIC 2018 Task 3 test set, with a predefined diagnosis distribution (AKIEC: 1, BCC: 1, BKL: 1, DF: 1, VASC: 1, MEL: 2, NV: 3). For each image a user had to choose one diagnosis out of seven options, and subsequently again after assistance from our foundation model, presented as multi-class probabilities visualized as bars and numbers for each class. Readers had the flexibility to complete multiple survey rounds with different image batches at their discretion, incompletely answered batches were omitted. The study was conducted online from August 20 to September 12, 2024, during which we collected data from 41 raters. Our foundation model for decision support utilized a weighted random sampler strategy, following the approach from⁶⁵ but excluding test-time augmentation. The model demonstrated robust performance, achieving an 80.4% mean (macro-averaged) recall, with notably high recall rates for critical skin lesions: 87.2% for melanoma and 86.0% for Basal Cell Carcinoma (BCC).

Evaluation metrics. For multi-class tasks, we primarily use a weighted F1 score, which averages class-specific F1 scores (harmonic means of precision and recall) weighted by class size. It addresses class imbalance in multi-class scenarios. For binary classification, we primarily use AUROC (Area Under the Receiver Operating Characteristic curve), measuring the model’s ability to distinguish between classes across all classification thresholds. An AUROC of 1.0 indicates perfect classification, while 0.5 suggests random guessing. This metric is particularly useful for imbalanced datasets and when we need to evaluate trade-offs between true positive and false positive rates. For the two reader studies, we report accuracy. In skin lesion segmentation, we use the Dice Similarity Coefficient (DSC) and Jaccard index (JAC) to assess segmentation quality. For TBP-based melanoma screening, we primarily report the sensitivity (recall) in malignant lesions, focusing on

the model’s ability to correctly identify malignant cases.

Statistical analysis. For skin tumor patch classification, melanoma slides classification, reader studies, metastasis prediction, and skin lesion segmentation, we conduct k-fold cross-validation due to either a relatively small sample size or following conventional practice. We compute the mean and standard deviation of performance across the folds, then calculate the standard error by dividing the standard deviation by the square root of the number of folds. The 95% confidence interval is derived using 1.96 times the standard error. To assess statistical significance, we conduct two-sided t-tests comparing PanDerm’s performance against the baseline model for each task. For the remaining datasets, we utilize nonparametric bootstrapping with 1,000 replicates to estimate 95% confidence intervals for each model’s performance. To compare models, we implement pairwise permutation tests, conducting 1,000 permutations per pair and recalculating performance metrics after each permutation. We derive two-sided P values to evaluate the null hypothesis that paired observations stem from identical distributions. Additionally, we perform t-tests to assess the statistical significance of inter-model performance variations. Our null hypothesis posits no discernible difference between PanDerm’s performance and that of its competitors. A P value < 0.05 was regarded as statistically significant.

Skin cancer and general skin condition classification datasets

HAM10000 ³⁵ (7 classes) : This dataset contains 10,015 dermoscopic images across 7 classes: actinic keratoses, basal cell carcinoma, benign keratosis, dermatofibroma, melanocytic nevi, melanoma, and vascular lesions. It is stratified into 60% training, 20% validation, and 20% test sets. For human-AI collaboration, we used the official dataset. All other experiments used the clean version from ⁸⁰, which prevents data leakage by ensuring lesions from the same patient are not split across sets.

BCN20000 ⁸¹ (9 classes) : This dataset comprises 12,413 dermoscopic images in 9 categories: nevus, melanoma, basal cell carcinoma, seborrheic keratosis, actinic keratosis, solar lentigo, squamous cell carcinoma, dermatofibroma, and vascular lesions, including lesions in hard-to-diagnose locations. It is similarly stratified (60/20/20 split). We used the clean version of BCN20000, which, like the HAM10000, addresses data leakage issues.

MSKCC ⁵⁷ (2 classes) : The dataset is curated from the MSKCC data from ISIC archive ⁵⁷, containing 8,984 dermoscopic images with melanoma and other classes.

HIBA ⁵⁷ (2 classes) : The dataset is curated from the HIBA data from ISIC archive ⁵⁷, containing 1,635 dermoscopic images with melanoma and other classes.

PAD-UFES-20 ⁴³ (6 classes) : This dataset from Brazil contains 2,298 close-up clinical images with 6 classes, including Actinic Keratosis, Basal Cell Carcinoma of skin, Malignant Melanoma, Melanocytic Nevus of Skin,

Squamous Cell Carcinoma, Seborrheic Keratosis.

DDI ⁶² (2 classes) : We grouped the classes of the DDI dataset into melanoma and others. The dataset contains 647 clinical images from the US.

Derm7pt ⁸² (2 classes) : Derm.D is a subset of Derm7pt, containing 839 dermoscopic images and Derm.C contains 839 clinical images with melanoma and other classes.

ISIC2024 ⁵⁰ (2 classes) : ISIC2024 is a multi-center dataset with skin lesion crops from total body photography (TBP). We chose a hold-out data with 49,025 crop images with three institutions (FNQH Cairns, Alfred Hospital, Melanoma Institute Australia) as the evaluation dataset.

PH2 ⁸³ (3 classes) : PH2 is a clinical image dataset from Portugal with 200 images and 3 classes. We reorganize it to a binary melanoma detection task.

Med-Node ⁸⁴ (2 classes) : The dataset contains 170 clinical images. We reorganize it to a binary melanoma detection task.

DermNet ⁴⁶ (23 classes) : DermNet contains 19,559 clinical images, the dataset consists of images of 23 types of skin diseases.

MMT-09 (9 classes) : The dataset is an in-house clinical dataset with 9 skin condition classes, including benign keratinocytic, malignant keratinocytic, melanocytic, inflammatory conditions and benign tumors, vascular lesion, basal cell carcinoma, malignant keratinocytic, melanoma, squamous cell carcinoma. We chose 38,476 images as our evaluation dataset.

MMT-74 (74 classes) : The dataset is an in-house clinical dataset with 74 general skin condition classes, which has more fine-grained classes built on the 9 classes of MMT-09. It includes general skin conditions like inflammatory, infective, benign proliferations, melanocytic, eczema. We chose the same 38,476 images as our evaluation dataset.

Skin Tumor Patch Classification (PATCH16) (16 classes) ⁷⁰ : The skin tumor patch classification task consists of tissue patches of 378 histopathology WSIs from the archive of the Institute of Pathology, Heidelberg University, the MVZ for Histology, Cytology and Molecular Diagnostics Trier and the Institute for Dermatopathology Hannover for classification of 16 categories including 4 tumor types and 12 normal tissue structures. We obtained a total of 129,364 image patches of $100 \times 100 \mu\text{m}$ (395×395) size. The dataset was stratified by label, with 55% allocated for training, 15% for validation, and 30% for testing.

Melanoma Slide Classification (WSI) (2 classes)⁸⁵ : The melanoma slide classification task from the National Cancer Institute’s Clinical Proteomic Tumor Analysis Consortium Cutaneous Melanoma (CPTAC-CM) cohort consists of histopathology WSIs for cancer detection. After selecting labeled WSIs, we obtained 302 slides (71 normal, 231 tumor). For training and evaluation, we employed a five-fold cross-validation strategy with label-stratified splits to maintain class balance.

Reader study1: Early melanoma detection based on SDDI-Alfred (2 classes) : The dataset consists of 179 serial dermoscopic imaging sequences from 122 patients, totaling 730 dermoscopic images. The patients were recruited from a private specialist dermatology clinic, with follow-up periods ranging from January 2007 to December 2019. Both melanoma and benign lesions that underwent short- or long-term sequential digital dermoscopic imaging (SDDI) at least once prior to biopsy were included. The dataset is well-balanced, with 90 benign lesions and 89 malignant lesions, encompassing both invasive and in situ melanomas. Of the 89 melanomas, 34 (38.2%) were invasive, with a mean Breslow thickness of 0.5 mm, while 55 (61.8%) were in situ. All lesions were monitored via digital dermoscopy, excised due to clinical concerns, and confirmed by pathological examination. The number of images per sequence varied from 1 to 12, with an average of approximately 4 images per sequence.

Longitudinal and melanoma metastasis datasets

Short-term lesion change detection based on SDDI1⁵⁷ (2 classes) : This dataset is sourced from the “repeated dermoscopic images of melanocytic lesions” by University Hospital Basel, available in the ISIC archive. It comprises 116 sequential lesions, each with a sequence length of 5, from 66 patients. The dataset is categorized into two classes for lesion change detection.

Short-term lesion change detection based on SDDI2 (2 classes) : SDDI2 is an in-house dataset from the Medical University of Vienna. It contains 229 sequential dermoscopic images with a sequence length of 2. The dataset includes both binary change labels and more fine-grained malignant change labels. This dataset is also used for short-term lesion change detection.

Melanoma metastasis prediction and prognosis based on CombinMel dataset (2 or 3 classes) : This dataset encompasses 680 dermoscopic images of invasive melanoma from 370 patients recruited across 10 hospital sites in multiple countries, including Australia and 5 European nations. For large melanomas, multiple images were captured to ensure comprehensive coverage of the entire lesion area. Among the 370 melanoma cases, the majority 261 (70.5%) are classified as Stage I, with 45 (12.2%) in Stage II, 61 (16.5%) in Stage III, and a small fraction of 3 (0.8%) in Stage IV. This distribution provides a comprehensive representation of melanoma stages, with a focus on earlier stages. Regarding metastasis status, 248 (67.0%) of cases showed no

metastasis, while 66 (17.8%) presented with metastasis at the time of diagnosis. Additionally, 56 (15.1%) of cases developed metastasis during the follow-up period.

Skin lesion segmentation based on ISIC2018 and HAM10000 : The skin lesion segmentation task is evaluated using two publicly available datasets. The ISIC2018 dataset⁵⁵ comprises 3,694 dermoscopic images with 2594 images for training, 100 for validation, and 1000 for testing. We follow this official dataset split for our experiments. The HAM10000 dataset³⁵ includes 10,015 dermoscopic images, each with corresponding binary segmentation labels. A randomized selection approach is adopted, with 64% of the images used for training, 16% for validation, and the remaining 20% for testing.

3D total body photography datasets

This dataset comprises 3D total body photography (TBP) images captured using the VECTRA WB360 system (Canfield Scientific Inc., Parsippany, NJ, USA). The system employs 92 cameras to simultaneously capture cross-polarised 2D images with standardized lighting within seconds, which are then merged to create a high-fidelity 3D avatar of each patient’s entire skin surface. From these 3D avatars, individual lesion tiles were exported for further analysis.

Photodamage risk assessment datasets (3 classes) : This in-house dataset⁸⁶ contains image tiles (693×693 pixels) created from 92 raw 2D photos, each representing approximately 10cm^2 of cutaneous surface. Tiles with $< 33\%$ skin surface were excluded using pixel color analysis. Manual review removed out-of-focus images, tiles with multiple body sites, or identifying features. The final dataset comprises 6,195 image tiles from MYM⁵³ and HOP⁵² studies, labeled as low, moderate, or severe photodamage risk labeled primarily by dermatology students.

Nevus counting datasets (2 classes) : This dataset, derived from the in-house MYM⁵³ study, contains 32,582 lesion tiles annotated as nevus or non-nevus. Three expert physicians independently labeled lesions on-screen, with consensus determined by ≥ 2 clinicians’ agreement. A senior dermatologist manually identified naevi in-clinic using a dermatoscope, serving as the gold standard for the test set. To ensure consistency, lesions under underwear, on the scalp, or on foot soles were excluded, and only lesions ≥ 2 mm were considered. A minimum one-month interval was maintained between on-screen and in-clinic labeling sessions.

Lesion risk prediction and TBP screening datasets (2 classes) : This dataset comprises 2,038 total body photography (TBP) scans from 480 patients, collected from the MYM and HOP studies. The raw TBP scans include nevi images and a variety of non-relevant images such as normal skin, scars, and freckles. To focus only on nevi, we applied filtering parameters based on built-in Vectra data settings: $\text{majorAxisMM} \geq 2$, $\text{deltaLBnorm} \geq 4.5$, $\text{out_of_bounds_fraction} \leq 0.25$, $\text{dnn_lesion_confidence} \geq 50$ and $\text{nevi_confidence} > 80$.

This process resulted in 196,933 lesion image tiles. We stratified the data by the patient for training, validation, and testing: 360 patients for training (146,752 images), 40 patients for validation (19,483 images), and 80 patients for testing (30,698 images, including 28 malignant lesions). Of the total dataset, 216 images represent malignant lesions, with 40 confirmed melanoma cases.

Measurements in TBP : Alongside the image tiles, Vectra provides a range of measurements for each lesion, mainly including size, color, and shape. Our TBP screening model incorporates 32 such measurements: “A”, “Aext”, “B”, “Bext”, “C”, “Cext”, “H”, “Hext”, “L”, “Lext”, “areaMM2”, “area_perim_ratio”, “color_std_mean”, “deltaA”, “deltaB”, “deltaL”, “deltaLB”, “deltaLBnorm”, “dnn_lesion_confidence”, “eccentricity”, “location_simple”, “majorAxisMM”, “minorAxisMM”, “nevi_confidence”, “norm_border”, “norm_color”, “perimeterMM”, “radial_color_std_max”, “stdL”, “stdLExt”, “symm_2axis”, and “symm_2axis_angle”.

Computing hardware and software

For self-supervised pretraining, we used $4 \times 80\text{GB}$ NVIDIA H100 GPUs configured for multi-GPU single-node training using DistributedDataParallel (DDP) as implemented by Python (v.3.9.13), PyTorch (v.2.2.1, CUDA 11.8) and Torchvision (v.0.17.1). The CAE-v2 code is used as the codebase to develop our foundation model, which can be found in its official repository¹. For downstream task evaluation, all experiments were conducted on $4 \times 49\text{GB}$ NVIDIA 6000 Ada GPUs. We used Python (v.3.9.19), PyTorch (v.2.2.2, CUDA 11.8), and Torchvision (v.0.17.2) for finetuning tasks, and Python (v.3.10.14), PyTorch (v.2.2.2, CUDA 11.8) and Torchvision (v.0.17.2) for linear probing tasks. We used Scikit-learn (v1.2.1) for logistic regression in the linear probing setting. Implementation of other comparative pretrained models was modified based on the official configuration in their respective repositories: MAE², SL_ImageNet³, DINOv2⁴, SwAVDerm⁵, autoSMIM⁶, BATFormer⁷, MedSAM⁸, ResNet50⁹, MILAN¹⁰, CLIP¹¹, BiomedCLIP¹², and MONET¹³.

Ethics statement

MYM study was approved by the Metro South Health Human Research Ethics Committee on 21 April 2016 (approval number: HREC/16/QPAH/125). Ethics approval has also been obtained from the University of Queensland Human Research Ethics Committee (approval number: 2016000554), Queensland University

¹github.com/Atten4Vis/CAE

²github.com/facebookresearch/mae

³huggingface.co/timm/vit_large_patch16_224.orig_in21k

⁴github.com/facebookresearch/dinov2

⁵github.com/shenyue-98/SwAVDerm

⁶github.com/Wzhjerry/autoSMIM

⁷github.com/xianlin7/BATFormer

⁸github.com/bowang-lab/MedSAM

⁹pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html

¹⁰github.com/zejiangh/MILAN

¹¹github.com/openai/CLIP

¹²huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224

¹³github.com/suinleelab/MONET/tree/main

of Technology Human Research Ethics Committee (approval number: 1600000515) and QIMR Berghofer (approval number: P2271). The HOP study has received Human Research Ethics Committee (HREC) approval from Metro South Health HREC (HREC/17/QPAH/816) and The University of Queensland HREC (2018000074). The ComBineMel dataset is part of the Computer biomarkers evaluation of invasive melanoma (ComBine Mel) study. The study was approved by the Alfred Hospital Ethics Committee on 08 August 2023 (approval number: HREC/98200/Alfred-2023). The study follows the National Statement on Ethical Conduct in Human Research (2007) protocols. SDDI2 dataset is approved by the Ethics Review Board of the Medical University of Vienna. MMT data study is part of a research agreement study with Monash eResearch Centre and was approved through the Monash University Human Research Ethics Committee (MUHREC). The NSSI dataset is part of the Brisbane Naevus Morphology Study, circa 2009-2014. The study followed the Declaration of Helsinki protocols and was approved by the Princess Alexandra Hospital human research ethics committee. The ACEMID_path study has received approval from the Alfred Hospital Ethics Committee (approval number: 746/23) to share data accrued for registered trial ACTRN12619001706167 (ACEMID) under the Metro South Human Research Committee protocol HREC/2019/QMS/57206 and the University of Queensland Human Research Ethics Committee protocol 2019003077. The SDDI_Alfred study has received approval from the Alfred Hospital Ethics Committee (approval number: 198/19) for use of sequential dermoscopic imaging data. Only de-identified retrospective data was used for research, without the active involvement of patients.

Data availability

Most datasets used in this study are publicly available. These datasets used for skin lesion diagnosis and segmentation tasks can be accessed through various repositories. The ISIC archive¹⁴ hosts several datasets, including MSKCC and HIBA. Other widely used benchmark datasets are available through their respective portals: BCN20000¹⁵, PAD-UFES-20¹⁶, DDI¹⁷, Derm7pt¹⁸, ISIC2024¹⁹, Med-Node²⁰, DermNet²¹, WSI²², PATCH16²³, ISIC2018_task1 and HAM1000²⁴, SDDI²⁵, and PH2²⁶. Access to in-house datasets is restricted due to patient privacy considerations. These include MMT for dermoscopic and clinical image pretraining and downstream multi-skin condition classification, NSSI for sequential dermoscopic image pretraining, ACEMID_path for dermatopathology pretraining, Edu1 and Edu2 for clinical image pretraining, SDDI2 for lesion change detection, SDDI_Alfred for reader study 1 (early-melanoma detection), and the TBP data from

¹⁴isic-archive.com

¹⁵figshare.com/articles/journal_contribution/BCN20000_Dermoscopic_Lesions_in_the_Wild/24140028/1

¹⁶kaggle.com/datasets/mahdavi1202/skin-cancer

¹⁷ddi-dataset.github.io/index.html

¹⁸derm.cs.sfu.ca/Welcome.html

¹⁹kaggle.com/competitions/isic-2024-challenge

²⁰kaggle.com/datasets/prabhavsanga/med-node

²¹kaggle.com/datasets/shubhamgoel27/dermnet

²²portal.gdc.cancer.gov/projects/TCGA-SKCM

²³heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/7QCR8S

²⁴challenge.isic-archive.com/data/

²⁵api.isic-archive.com/collections/328/

²⁶fc.up.pt/addi/ph2%20database.html

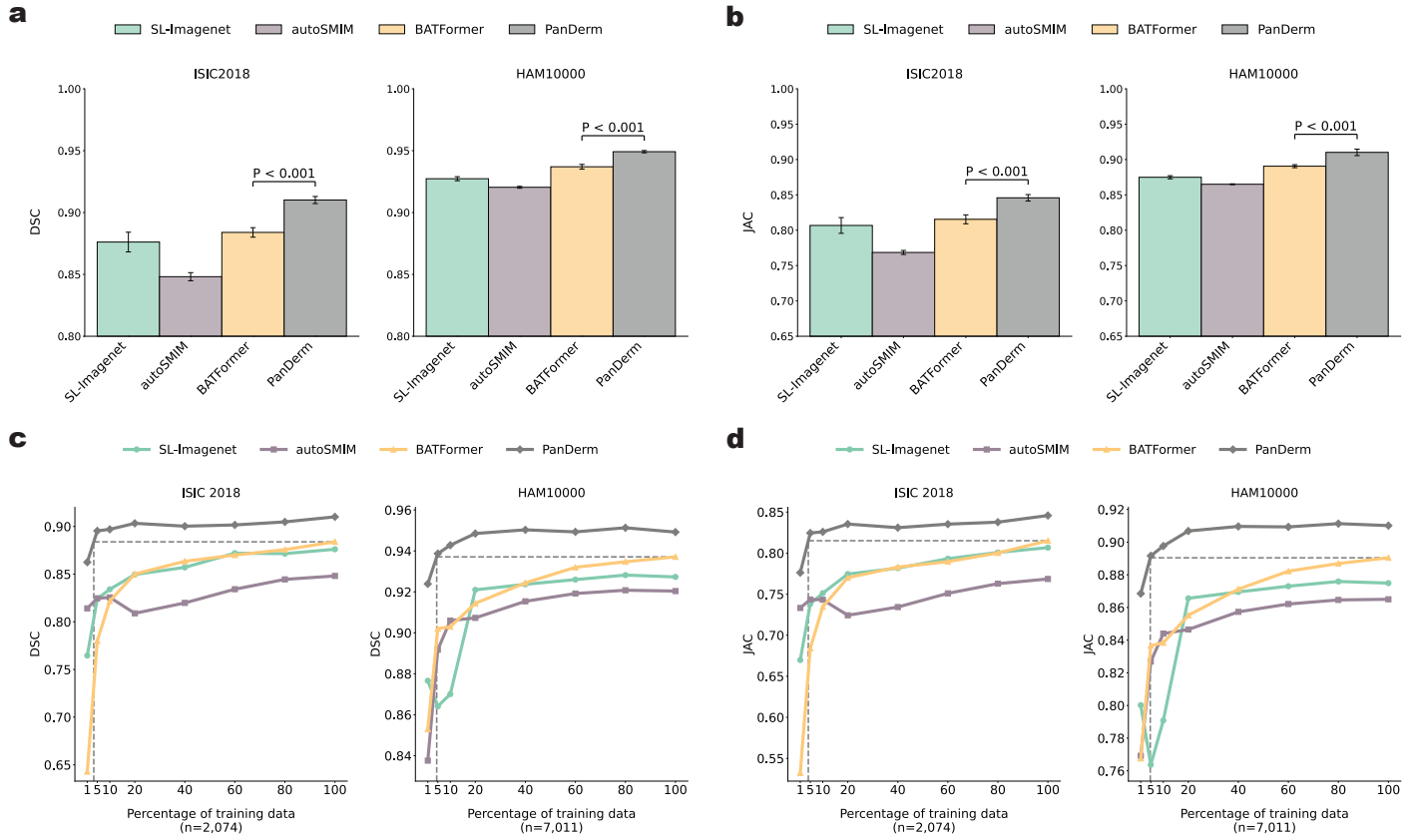
MYM and HOP studies for all TBP-based pretraining and evaluation. Researchers interested in accessing these datasets should direct their requests to the corresponding author. Requests will be evaluated according to institutional and departmental policies to ensure compliance with intellectual property rights and patient privacy obligations. The availability of these data may be subject to additional restrictions or requirements.

Code availability

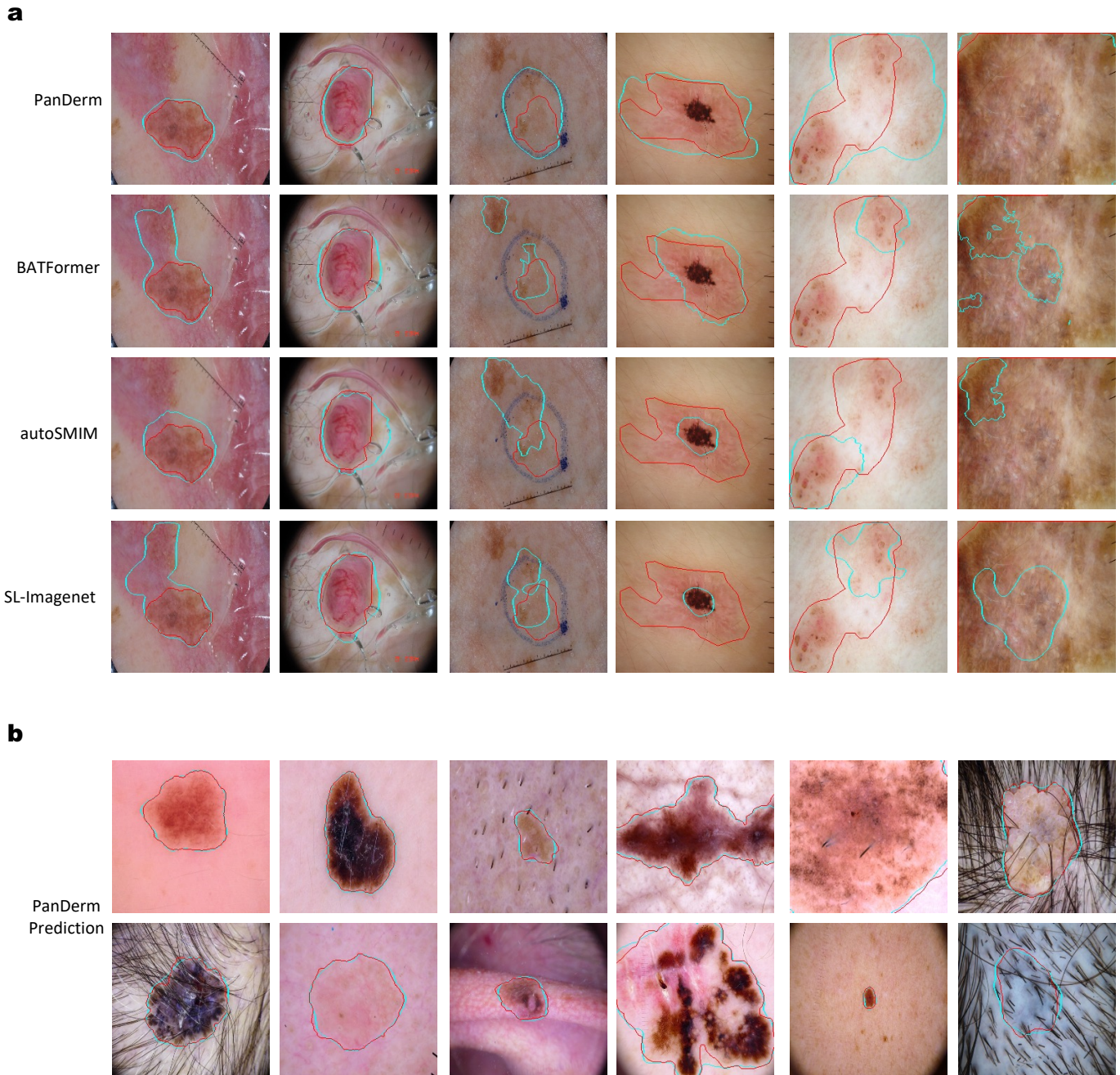
We have made the encoder code and weights available for downstream task applications. They can be accessed at <https://github.com/SiyuanYan1/PanDerm>. We have documented all experiments in detail in our Methods section to enable independent replication. To facilitate the broader use of our model, we have provided tutorial Jupyter notebooks and downstream evaluation code suitable for a wide scientific audience. These resources have been made available to ensure transparency and to promote further research in this field.

Author contributions

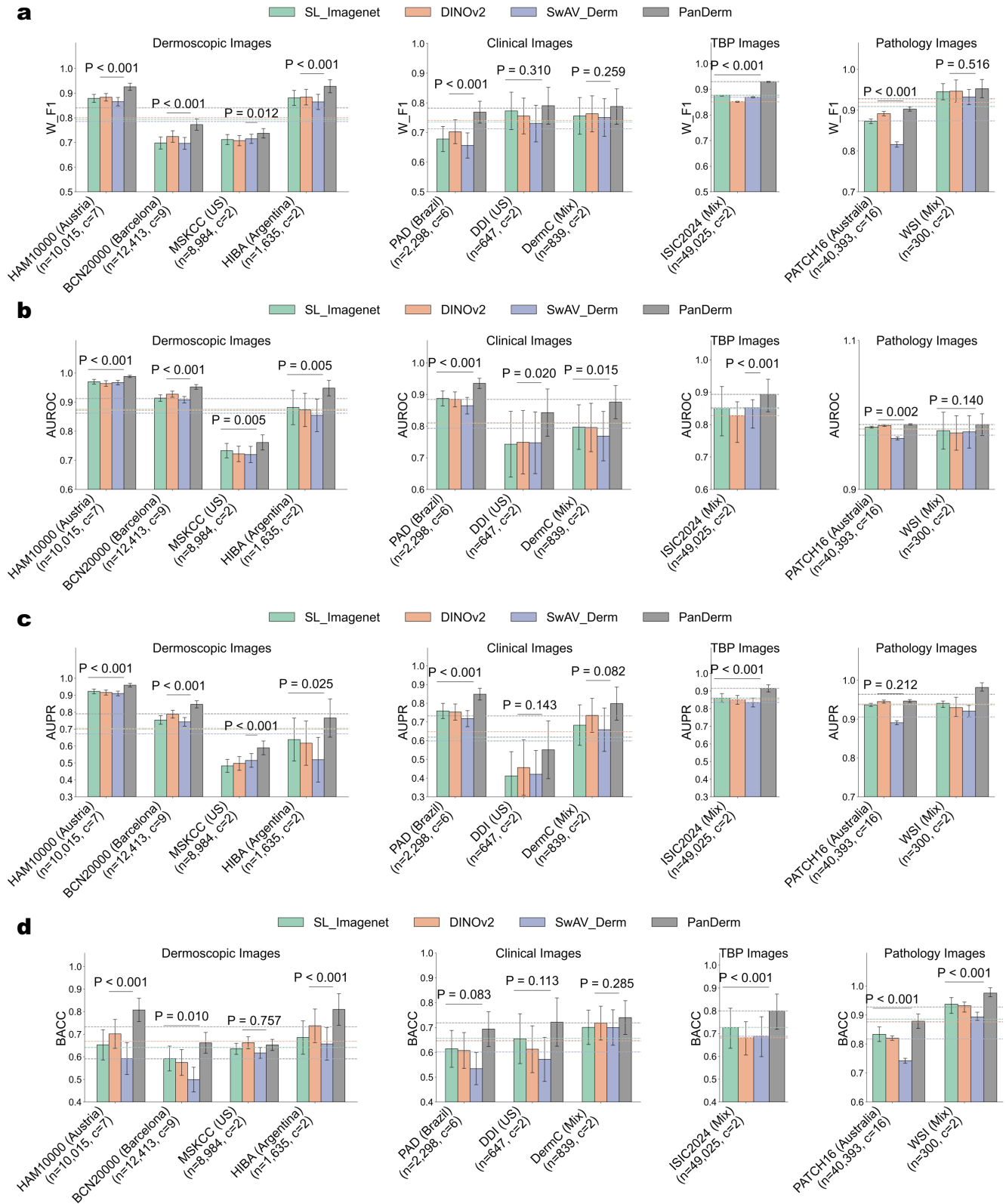
S.Y., Z.Y., H.K., H.P.S., and Z.G. conceived the study and designed experiments. S.Y., C.V.A., M.H., L.Y., H.K., V.M., M.J., H.P.S., and Z.G. contributed to data acquisition, preprocessing, and organization. S.Y. performed model development and pretraining. S.Y., Z.Y., Z.W., L.Y., H.K., and P.T. contributed to downstream task evaluation. S.Y., Z.Y., C.V.A., and V.M. performed experimental analyses regarding metastasis prediction and prognosis. S.Y. and Z.Y. performed experimental analyses regarding TBP-based screening. Z.W. and S.Y. performed experimental analysis regarding lesion segmentation. S.Y. and L.Y. performed experimental analysis regarding dermatopathology image analysis. S.Y., H.K., and P.T. contributed to the human-AI collaboration reader study experimental analysis. S.Y. also contributed to all remaining tasks of experimental analysis. H.K. and P.T. developed the web-based reader platforms and conducted the human-AI reader study. V.M. and Z.Y. contributed to the early melanoma detection reader study data. H.K., P.T., C.V.A., C.P., V.M., M.J., and H.P.S. contributed clinical inputs to this research. G.T., V.T., A.B.N., D.P., P.B., and S.S. provided computing resources and data management. All authors contributed to the drafting and revising of the manuscript.



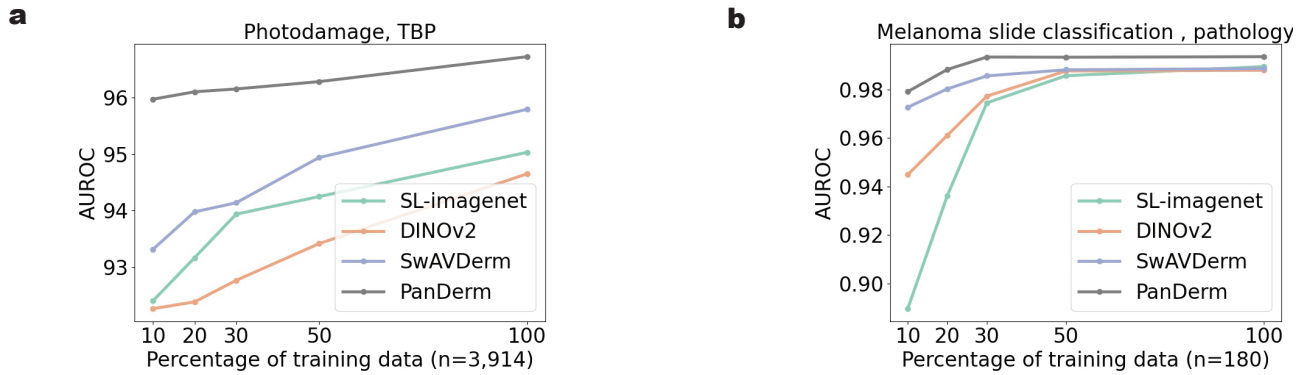
Extended Data Figure 1: **Quantitative skin lesion segmentation results.** **a, b.** Segmentation performance measured by dice score (DSC) and Jaccard index (JAC) for PanDerm and baseline models on ISIC2018 and HAM10000 datasets. **c, d.** Label efficiency generalization performance for PanDerm and baselines, showing mean DSC and JAC on ISIC2018 and HAM10000 datasets. Error bars in **a, b** indicate 95% confidence intervals; bar centers represent mean values. Points in **c, d** denote mean values. All estimates are derived from five replicas with different seeds. Statistical significance was assessed using two-sided t-tests.



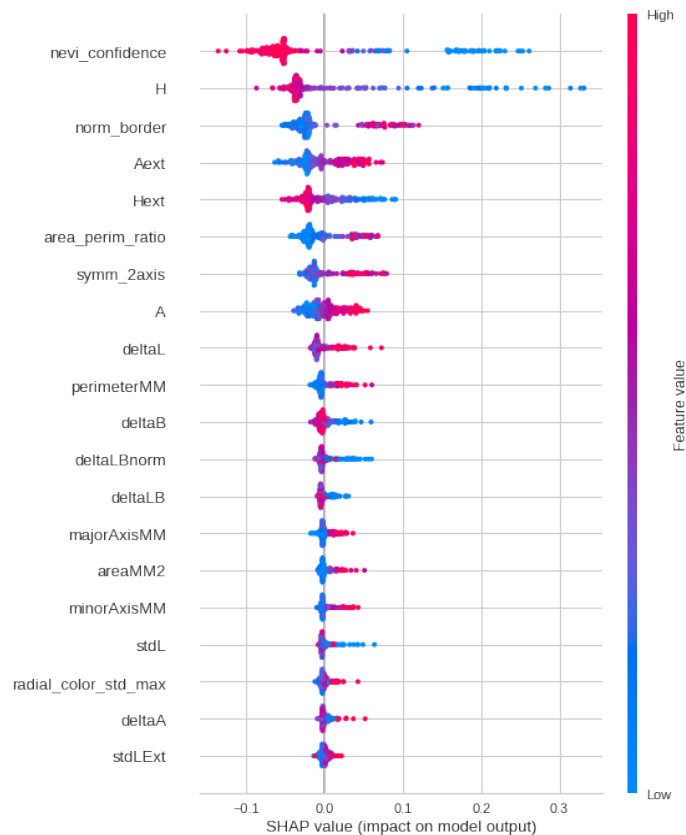
Extended Data Figure 2: **Qualitative skin lesion segmentation results.** **a.** Comparison of PanDerm against baseline models on challenging examples from HAM10000. Red contours indicate ground truth masks, while cyan contours show model predictions. **b.** PanDerm segmentation results on a random selection of images from HAM10000.



Extended Data Figure 3: Performance of PanDerm versus other pretrained models on 10 pigmented skin lesion datasets across multiple centers and modalities. a. Performances are measured by weighted F1 (W_F1). **b.** Performances are measured by AUROC. **c.** Performances are measured by AUPR. **d.** Performances are measured by BACC. n: data size, c: class number. Dashed lines show the average performance of each model across different datasets.



Extended Data Figure 4: **Label efficiency generalization results on additional tasks.** **a.** Label efficiency analysis for photodamage risk assessment using Total Body Photography (TBP) images. Results demonstrate model performance with limited labeled data available. PanDerm outperformed the second-best models using only 10% of labeled images. **b.** Label efficiency analysis for melanoma classification using whole slide dermatopathology images. Results illustrate model performance with limited labeled data. PanDerm surpassed the second-best models using less than 30% of labeled images.



Extended Data Figure 5: **SHAP (SHapley Additive exPlanations) value plot.** It shows the impact of various measurement variables captured by the 3D TBP machine on the model output. The plot displays the relative importance and directional influence of each feature, with colors indicating high (red) to low (blue) feature values, and the x-axis representing the SHAP value or impact on the model's prediction. Features are ordered by their overall importance, with 'nevi_confidence' having the highest impact and 'stdLExt' the lowest.

Dataset	Model	W_F1	AUROC	BACC	AUPR
PAD	SL_Imagenet	0.678 (0.636-0.720)***	0.887 (0.864-0.911)***	0.614 (0.540-0.688)	0.759 (0.718-0.799)***
	DINOV2	0.702 (0.662-0.743)**	0.885 (0.861-0.908)***	0.607 (0.535-0.679)*	0.753 (0.710-0.796)***
	SwaVDerm	0.656 (0.614-0.698)***	0.865 (0.838-0.891)***	0.534 (0.469-0.599)**	0.718 (0.675-0.761)***
	PanDerm	0.768 (0.732-0.805)	0.935 (0.919-0.951)	0.694 (0.624-0.764)	0.849 (0.817-0.880)
HAM10000	SL_Imagenet	0.879 (0.863-0.895)***	0.970 (0.962-0.978)***	0.653 (0.586-0.720)***	0.922 (0.909-0.936)***
	DINOV2	0.883 (0.868-0.899)***	0.964 (0.954-0.974)***	0.701 (0.637-0.765)***	0.916 (0.901-0.931)***
	SwaVDerm	0.865 (0.848-0.883)***	0.967 (0.959-0.975)***	0.592 (0.521-0.663)***	0.910 (0.896-0.924)***
	PanDerm	0.926 (0.912-0.940)	0.988 (0.984-0.992)	0.807 (0.756-0.859)	0.959 (0.949-0.970)
DermC	SL_Imagenet	0.756 (0.694-0.818)	0.797 (0.726-0.867)*	0.700 (0.631-0.770)	0.683 (0.575-0.791)*
	DINOV2	0.763 (0.702-0.824)	0.796 (0.719-0.872)**	0.717 (0.649-0.786)	0.735 (0.644-0.826)
	SwaVDerm	0.750 (0.687-0.814)	0.768 (0.690-0.846)**	0.700 (0.628-0.771)	0.658 (0.543-0.774)**
	PanDerm	0.788 (0.728-0.847)	0.876 (0.824-0.928)	0.740 (0.672-0.808)	0.798 (0.710-0.886)
BCN20000	SL_Imagenet	0.698 (0.673-0.722)***	0.914 (0.903-0.925)***	0.592 (0.537-0.647)*	0.754 (0.728-0.779)***
	DINOV2	0.724 (0.701-0.747)***	0.927 (0.917-0.938)***	0.575 (0.518-0.632)**	0.787 (0.765-0.810)***
	SwaVDerm	0.696 (0.672-0.720)***	0.908 (0.897-0.919)***	0.499 (0.444-0.554)***	0.742 (0.717-0.768)***
	PanDerm	0.772 (0.750-0.795)	0.952 (0.944-0.960)	0.662 (0.616-0.708)	0.846 (0.825-0.867)
DDI	SL_Imagenet	0.773 (0.710-0.836)	0.743 (0.639-0.847)**	0.655 (0.554-0.755)	0.412 (0.283-0.541)*
	DINOV2	0.756 (0.695-0.816)	0.749 (0.649-0.849)*	0.612 (0.518-0.706)*	0.456 (0.308-0.605)
	SwaVDerm	0.730 (0.668-0.792)	0.747 (0.650-0.845)**	0.571 (0.483-0.660)**	0.421 (0.294-0.548)*
	PanDerm	0.790 (0.728-0.852)	0.843 (0.768-0.918)	0.722 (0.624-0.819)	0.551 (0.397-0.705)
HIBA	SL_Imagenet	0.881 (0.850-0.911)**	0.881 (0.823-0.940)**	0.685 (0.612-0.759)**	0.638 (0.511-0.765)*
	DINOV2	0.884 (0.852-0.915)***	0.873 (0.816-0.930)**	0.737 (0.663-0.812)***	0.616 (0.486-0.747)*
	SwaVDerm	0.865 (0.834-0.895)***	0.854 (0.799-0.910)***	0.657 (0.585-0.729)***	0.519 (0.387-0.651)***
	PanDerm	0.928 (0.901-0.954)	0.948 (0.922-0.975)	0.810 (0.740-0.880)	0.765 (0.652-0.878)
MSKCC	SL_Imagenet	0.712 (0.691-0.732)**	0.733 (0.708-0.758)**	0.635 (0.611-0.660)	0.482 (0.444-0.521)***
	DINOV2	0.707 (0.687-0.728)**	0.722 (0.695-0.748)**	0.662 (0.636-0.689)	0.497 (0.457-0.537)***
	SwaVDerm	0.715 (0.696-0.733)*	0.720 (0.692-0.748)***	0.617 (0.593-0.641)**	0.515 (0.474-0.556)***
	PanDerm	0.737 (0.718-0.756)	0.761 (0.735-0.787)	0.653 (0.628-0.677)	0.589 (0.548-0.630)
PATCH16	SL_Imagenet	0.873 (0.867-0.878)***	0.992 (0.991-0.992)***	0.834 (0.808-0.859)***	0.936 (0.932-0.941)***
	DINOV2	0.892 (0.886-0.897)***	0.993 (0.992-0.994)**	0.820 (0.813-0.828)***	0.945 (0.941-0.949)
	SwaVDerm	0.816 (0.809-0.822)***	0.984 (0.983-0.985)***	0.742 (0.734-0.751)***	0.891 (0.885-0.896)***
	PanDerm	0.903 (0.898-0.908)	0.994 (0.993-0.994)	0.879 (0.854-0.903)	0.946 (0.943-0.950)
ISIC2024	SL_Imagenet	0.877 (0.873-0.877)***	0.849 (0.765-0.917)***	0.727 (0.635-0.811)**	0.860 (0.835-0.885)***
	DINOV2	0.851 (0.849-0.853)***	0.827 (0.745-0.870)***	0.682 (0.606-0.752)***	0.850 (0.825-0.875)***
	SwaVDerm	0.869 (0.866-0.871)***	0.852 (0.789-0.877)***	0.689 (0.599-0.774)***	0.835 (0.810-0.860)***
	PanDerm	0.929 (0.927-0.931)	0.893 (0.839-0.940)	0.799 (0.718-0.873)	0.915 (0.895-0.935)
WSI	SL_Imagenet	0.945 (0.925-0.965)	0.989 (0.977-1.002)	0.937 (0.906-0.960)***	0.941 (0.930-0.947)***
	DINOV2	0.947 (0.919-0.974)	0.988 (0.976-1.000)	0.932 (0.910-0.945)***	0.930 (0.906-0.956)***
	SwaVDerm	0.932 (0.914-0.951)	0.989 (0.978-1.000)*	0.893 (0.882-0.910)***	0.920 (0.905-0.935)***
	PanDerm	0.953 (0.930-0.975)	0.994 (0.986-1.001)	0.976 (0.963-0.994)	0.981 (0.972-0.993)

Extended Data Table 1: **Skin cancer diagnosis performance of different models across multinational datasets.** Models include SL_Imagenet (supervised learning on ImageNet), DINOV2, SwaVDerm, and PanDerm. Performance is reported using Weighted F1 score (W_F1), Area Under the Receiver Operating Characteristic curve (AUROC), Balanced Accuracy (BACC), and Area Under the Precision-Recall curve (AUPR). Further details on the experimental setup, datasets, and metrics are provided in **Methods**. Best-performing model for each metric and dataset is bolded and highlighted. 95% CI is included in parentheses. Significance levels for comparisons with the best model: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Model	Dataset	W_F1	AUROC	BACC	AUPR
SL_Imagenet	MMT-09	0.661 (0.653, 0.668)***	0.860 (0.846, 0.874)***	0.404 (0.377, 0.430)***	0.482 (0.460, 0.503)***
DINOv2	MMT-09	0.672 (0.657, 0.687)***	0.858 (0.839, 0.877)***	0.433 (0.412, 0.454)***	0.474 (0.459, 0.490)***
SwaVDerm	MMT-09	0.624 (0.615, 0.634)***	0.814 (0.802, 0.825)***	0.356 (0.347, 0.365)***	0.411 (0.401, 0.422)***
PanDerm	MMT-09	0.704 (0.699, 0.709)	0.901 (0.888, 0.913)	0.462 (0.436, 0.488)	0.560 (0.539, 0.581)
SL_Imagenet	MMT-74	0.417 (0.411, 0.423)***	0.822 (0.807, 0.836)***	0.119 (0.101, 0.138)***	0.144 (0.125, 0.162)***
DINOv2	MMT-74	0.414 (0.401, 0.428)***	0.842 (0.830, 0.853)***	0.115 (0.103, 0.127)***	0.146 (0.127, 0.165)***
SwaVDerm	MMT-74	0.349 (0.340, 0.358)***	0.774 (0.758, 0.790)***	0.085 (0.073, 0.098)***	0.105 (0.085, 0.124)***
PanDerm	MMT-74	0.488 (0.482, 0.494)	0.887 (0.872, 0.902)	0.174 (0.159, 0.189)	0.211 (0.186, 0.235)
SL_Imagenet	DermNet	0.497 (0.481, 0.512)***	0.885 (0.878, 0.892)***	0.462 (0.444, 0.480)***	0.426 (0.407, 0.444)***
DINOv2	DermNet	0.536 (0.521, 0.551)***	0.902 (0.896, 0.909)***	0.505 (0.487, 0.523)***	0.476 (0.456, 0.496)***
SwaVDerm	DermNet	0.474 (0.458, 0.490)***	0.884 (0.878, 0.891)***	0.442 (0.424, 0.460)***	0.428 (0.410, 0.446)***
PanDerm	DermNet	0.619 (0.603, 0.634)	0.944 (0.939, 0.949)	0.586 (0.568, 0.603)	0.623 (0.603, 0.642)

Extended Data Table 2: **General multi-class skin condition classification performance of different models on MMT-09, MMT-74, and Dermnet datasets.** All models were evaluated on three datasets: MMT-09, MMT-74, and DermNet. Models include SL_Imagenet (supervised learning on ImageNet), DINOv2, SwaVDerm, and PanDerm. Performance is reported using Weighted F1 score (W_F1), Area Under the Receiver Operating Characteristic curve (AUROC), Balanced Accuracy (BACC), and Area Under the Precision-Recall curve (AUPR). The best-performing model for each metric and dataset is bolded. 95% CI is included in parentheses. *** $p < 0.001$ compared to PanDerm.

Dataset	Model	AUROC	Sensitivity	Specificity	BACC
DDI1P	Default	0.596 (0.567-0.624)***	0.173 (0.141-0.205)***	0.969 (0.939-0.988)	0.571 (0.542-0.611)***
	w/ Warp	0.673 (0.643-0.702)***	0.533 (0.493-0.562)***	0.765 (0.735-0.794)***	0.649 (0.608-0.685)***
	w/ Mask	0.648 (0.629-0.662)***	0.600 (0.571-0.626)***	0.646 (0.617-0.675)***	0.623 (0.594-0.652)***
	w/ Whole pipeline	0.706 (0.686-0.725)	0.653 (0.634-0.673)	0.741 (0.722-0.751)***	0.697 (0.688-0.717)
DDI2P	Default	0.683 (0.517-0.849)***	0.940 (0.864-1.000)	0.239 (0.115-0.593)***	0.590 (0.449-0.730)***
	w/ Warp	0.710 (0.579-0.841)***	0.942 (0.862-1.000)	0.273 (0.013-0.533)***	0.607 (0.506-0.709)***
	w/ Mask	0.695 (0.564-0.822)***	0.935 (0.853-0.995)	0.255 (0.012-0.511)***	0.595 (0.495-0.695)***
	w/ Whole pipeline	0.767 (0.649-0.886)	0.854 (0.797-0.911)***	0.577 (0.387-0.768)	0.716 (0.621-0.810)

Extended Data Table 3: **Ablation study on pre-processing methods for short-term lesion change detection based on SDDI1P and SDDI2P datasets.** Metrics: AUROC, Sensitivity, Specificity, BACC (Balanced Accuracy). Warp denoted image registration, Mask denoted lesion segmentation, and the Whole pipeline denoted our proposed pre-processing pipeline. The best model is bolded and highlighted. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Dataset	Model	AUROC	Sensitivity	Specificity	BACC
SDDI1	SL_Imagenet	0.616 (0.599-0.634)***	0.520 (0.501-0.543)***	0.647 (0.628-0.669)***	0.584 (0.567-0.613)***
	DINOV2	0.660 (0.649-0.678)***	0.573 (0.554-0.592)***	0.601 (0.586-0.622)***	0.587 (0.559-0.607)***
	SwaVDerm	0.632 (0.614-0.652)***	0.191 (0.163-0.214)***	0.985 (0.961-0.999)	0.588 (0.567-0.604)***
	PanDerm	0.706 (0.686-0.725)	0.653 (0.634-0.673)	0.741 (0.722-0.751)***	0.697 (0.688-0.717)
SDDI2	SL_Imagenet	0.715 (0.594-0.837)	0.870 (0.737-1.000)	0.392 (0.234-0.550)	0.631 (0.582-0.681)
	DINOV2	0.730 (0.533-0.928)	0.826 (0.673-0.979)	0.584 (0.160-1.000)	0.705 (0.556-0.853)
	SwaVDerm	0.656 (0.547-0.764)*	0.970 (0.920-1.000)	0.181 (0.051-0.412)**	0.575 (0.482-0.669)*
	PanDerm	0.767 (0.649-0.886)	0.854 (0.797-0.911)	0.577 (0.387-0.768)	0.716 (0.621-0.810)

Extended Data Table 4: **Short-term lesion change detection performance of different models on SDDI1 and SDDI2 datasets.** Models: SL_Imagenet, DINOV2, SwaVDerm, and PanDerm. Metrics: AUROC, Sensitivity, Specificity, BACC (Balanced Accuracy). The best model is bolded and highlighted. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Dataset	Model	AUROC	BACC
SDDI2M	SL_Imagenet	0.690 (0.580-0.800)**	0.588 (0.435-0.740)***
	DINOV2	0.665 (0.565-0.755)**	0.614 (0.460-0.770)***
	SwaVDerm	0.650 (0.558-0.742)**	0.540 (0.429-0.762)***
	PanDerm	0.840 (0.769-0.911)	0.660 (0.472-0.848)

Extended Data Table 5: **Malignant lesion change detection performance of different models on SDDI2 dataset.** classes: malignant lesion change vs others. Models: SL_Imagenet, DINOV2, SwaVDerm, and PanDerm. Metrics: AUROC and BACC (Balanced Accuracy). The best model is bolded and highlighted. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Dataset	Model	W_F1	AUROC	BACC	AUPR
Combinemel (2 classes)	SL_Imagenet	0.839 (0.781-0.897)	0.938 (0.898-0.978)	0.833 (0.767-0.899)	0.915 (0.867-0.964)
	DINOV2	0.841 (0.780-0.901)*	0.921 (0.872-0.971)**	0.842 (0.778-0.907)	0.886 (0.823-0.949)***
	SwaVDerm	0.847 (0.790-0.904)	0.944 (0.905-0.983)	0.858 (0.796-0.920)	0.924 (0.876-0.962)
	PanDerm	0.889 (0.837-0.941)	0.964 (0.937-0.991)	0.882 (0.822-0.942)	0.944 (0.907-0.982)
Combinemel (3 classes)	SL_Imagenet	0.661 (0.598-0.725)*	0.834 (0.789-0.879)**	0.526 (0.429-0.624)	0.734 (0.694-0.773)*
	DINOV2	0.632 (0.571-0.692)*	0.833 (0.785-0.880)**	0.474 (0.376-0.571)**	0.728 (0.689-0.767)**
	SwaVDerm	0.693 (0.633-0.744)*	0.875 (0.832-0.917)*	0.601 (0.505-0.697)	0.775 (0.728-0.822)
	PanDerm	0.721 (0.662-0.780)	0.896 (0.860-0.932)	0.624 (0.530-0.719)	0.792 (0.746-0.838)

Extended Data Table 6: **Metastasis prediction performance of different models on Combinemel dataset (2 classes and 3 classes).** 2 classes: metastasis vs control (no metastasis). 3 classes: local metastasis vs distant metastasis vs control. Models include SL_Imagenet (supervised learning on ImageNet), DINOV2, SwaVDerm, and PanDerm. Metrics: Weighted F1 score (W_F1), Area Under the Receiver Operating Characteristic curve (AUROC), Balanced Accuracy (BACC), and Area Under the Precision-Recall curve (AUPR). The best-performing model for each metric and dataset is bolded and highlighted. 95% CI in parentheses. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Dataset	Model	W_F1	AUROC	BACC	AUPR
Solardamage	SL_Imagenet	0.890 (0.873-0.907)	0.950 (0.939-0.962)	0.829 (0.799-0.860)	0.899 (0.874-0.924)
	DINOV2	0.868 (0.849-0.888)	0.947 (0.935-0.958)	0.810 (0.779-0.840)	0.875 (0.847-0.903)
	SwaVDerm	0.888 (0.870-0.905)	0.958 (0.948-0.968)	0.826 (0.795-0.857)	0.908 (0.886-0.931)
	PanDerm	0.896 (0.879-0.913)	0.961 (0.951-0.971)	0.845 (0.816-0.874)	0.917 (0.893-0.940)

Extended Data Table 7: **Solar damage risk assessment performance of different models on HOP&MYM_solar dataset.** 3 classes: low vs medium vs high risk. Models include SL_Imagenet (supervised learning on ImageNet), DINOV2, SwaVDerm, and PanDerm. Metrics: Weighted F1 score (W_F1), Area Under the Receiver Operating Characteristic curve (AUROC), Balanced Accuracy (BACC), and Area Under the Precision-Recall curve (AUPR). The best-performing model for each metric is bolded and highlighted. 95% CI in parentheses.

Dataset	Model	W_F1	AUROC	BACC	AUPR
MYM	SL_Imagenet	0.952 (0.950-0.954)***	0.979 (0.973-0.985)***	0.810 (0.803-0.817)***	0.836 (0.821-0.850)***
	DINOV2	0.956 (0.953-0.959)	0.980 (0.977-0.983)***	0.838 (0.825-0.851)***	0.834 (0.808-0.861)***
	SwaVDerm	0.944 (0.938-0.950)***	0.973 (0.969-0.978)***	0.774 (0.765-0.782)***	0.819 (0.799-0.839)***
	PanDerm	0.956 (0.953-0.958)	0.983 (0.979-0.987)	0.820 (0.799-0.842)	0.844 (0.820-0.868)

Extended Data Table 8: **Nevus counting performance of different models on a subset of MYM dataset.** Models include SL_Imagenet (supervised learning on ImageNet), DINOV2, SwaVDerm, and PanDerm. Metrics: Weighted F1 score (W_F1), Area Under the Receiver Operating Characteristic curve (AUROC), Balanced Accuracy (BACC), and Area Under the Precision-Recall curve (AUPR). The best-performing model for each metric is bolded and highlighted. 95% CI in parentheses. *** $p < 0.001$ compared to PanDerm.

Model	AUC	W-F1	BACC
SL-Imagenet	68.09 (67.40-68.80)***	71.78 (71.27-72.33)***	62.09 (61.56-62.59)***
DINOV2	67.83 (66.69-68.03)***	69.83 (68.92-70.34)***	59.40 (58.75-59.92)***
SwavDerm	66.74 (66.22-67.10)***	68.15 (67.76-68.59)***	61.01 (60.53-61.49)***
PanDerm	70.47 (69.76-71.15)	73.63 (62.74-73.92)	65.72 (65.13-66.28)

Extended Data Table 9: **Risk stratification performance of different models on HOP&MYM datasets.** The table shows the performance metrics for different models. Metrics include Area Under the Curve (AUC), Weighted F1 score (W-F1), and Balanced Accuracy (BACC). The best-performing model for each metric is bolded and highlighted. 95% CI in parentheses. All p-values < 0.001 (***).

Model	Prediction head	Benign			Malignant		
		Precision	Recall	F1-score	Precision	Recall	F1-score
SL-Imagenet	UD	0.999	0.943	0.971	0.007	0.464	0.015
	CLS	1.000	0.955	0.977	0.013	0.643	0.025
	CMB	1.000	0.921	0.959	0.008	0.679	0.015
	CMB_ML	1.000	0.883	0.938	0.006	0.821	0.013
DINOv2	UD	1.000	0.967	0.983	0.014	0.500	0.027
	CLS	1.000	0.954	0.977	0.013	0.679	0.026
	CMB	1.000	0.939	0.968	0.010	0.679	0.020
	CMB_ML	1.000	0.900	0.947	0.007	0.821	0.015
SwavDerm	UD	0.999	0.958	0.978	0.009	0.393	0.017
	CLS	1.000	0.961	0.980	0.012	0.500	0.023
	CMB	1.000	0.936	0.967	0.008	0.571	0.016
	CMB_ML	1.000	0.896	0.945	0.007	0.857	0.015
PanDerm	UD	1.000	0.943	0.971	0.009	0.571	0.018
	CLS	1.000	0.971	0.985	0.016	0.500	0.031
	CMB	1.000	0.928	0.962	0.009	0.714	0.018
	CMB_ML	1.000	0.887	0.940	0.007	0.893	0.014

Extended Data Table 10: **TBP-based Malignant lesion screening performance of different models and prediction head types on HOP&MYM dataset.** Malignant recall is the most crucial metrics. Results are shown for both benign and malignant classifications, including precision, recall, and F1-score.

Dataset	Model	DSC	JAC
ISIC2018	SL-Imagenet	0.876 (0.870-0.887)***	0.807 (0.799-0.822)***
	autoSMIM	0.848 (0.845-0.851)***	0.769 (0.766-0.771)***
	BATFormer	0.884 (0.880-0.889)***	0.815 (0.809-0.823)***
	PanDerm	0.910 (0.907-0.913)	0.846 (0.842-0.850)

Extended Data Table 11: **Lesion segmentation performance of different models on ISIC2018 dataset.** Models include SL-Imagenet, autoSMIM, BATFormer, and PanDerm. Metrics: Dice Similarity Coefficient (DSC) and Jaccard Index (JAC). The best-performing model for each metric is bolded and highlighted. 95% CI in parentheses. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Dataset	Model	DSC	JAC
HAM10000	SL-Imagenet	0.927 (0.926-0.929)***	0.875 (0.873-0.878)***
	autoSMIM	0.920 (0.920-0.921)***	0.865 (0.864-0.866)***
	BATFormer	0.937 (0.935-0.939)***	0.891 (0.889-0.893)***
	PanDerm	0.949 (0.949-0.950)	0.910 (0.908-0.917)

Extended Data Table 12: **Lesion segmentation performance of different models on HAM10000 dataset.** Models include SL-Imagenet, autoSMIM, BATFormer, and PanDerm. Metrics: Dice Similarity Coefficient (DSC) and Jaccard Index (JAC). The best-performing model for each metric is bolded and highlighted. 95% CI in parentheses. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Dataset	Model	DSC	JAC
ISIC2018	MedSAM	0.904 (0.900-0.911)	0.841 (0.836-0.848)
	PanDerm	0.910 (0.907-0.913)*	0.846 (0.842-0.850)*
HAM10000	MedSAM	0.949 (0.948-0.951)	0.905 (0.904-0.907)
	PanDerm	0.949 (0.949-0.950)	0.910 (0.908-0.917)

Extended Data Table 13: **Lesion segmentation performance comparison of PanDerm and MedSAM on ISIC2018 and HAM10000 datasets.** Metrics: Dice Similarity Coefficient (DSC) and Jaccard Index (JAC). The best-performing model for each metric is bolded and highlighted. 95% CI in parentheses. Significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. For ISIC2018: DSC p-value = 0.017, JAC p-value = 0.025. For HAM10000: DSC p-value = 0.793, JAC p-value = 0.112.

GPU	Model	Dataset	Training Time	Inference Time
A6000	PanDerm	ISIC2018	2h 11min	52s
		HAM10000	6h 29min	59s
	MedSAM	ISIC2018	8h 32min	2m 8s
		HAM10000	28h 57min	4m 4s
RTX3090	PanDerm	ISIC2018	2h 33min	46s
		HAM10000	7h 48min	1m 6s
	MedSAM	ISIC2018	11h 58min	2m 16s
		HAM10000	38h 23min	4m 22s

Extended Data Table 14: **Training and inference times comparison between PanDerm and MedSAM on lesion segmentation.** The table shows the training and inference times for PanDerm and MedSAM on ISIC2018 and HAM10000 datasets, using A6000 and RTX3090 GPUs.

Percent	Model	W_F1	AUROC	BACC	AUPR
5%	SL_ImageNet	0.792 (0.774-0.809)***	0.919 (0.905-0.934)***	0.337 (0.294-0.379)***	0.853 (0.838-0.867)***
	DINOv2	0.803 (0.786-0.821)***	0.928 (0.913-0.943)***	0.380 (0.322-0.437)***	0.852 (0.837-0.867)***
	SwAVDerm	0.793 (0.776-0.810)***	0.917 (0.901-0.932)***	0.322 (0.288-0.356)***	0.841 (0.826-0.855)***
	PanDerm	0.851 (0.834-0.868)	0.960 (0.950-0.970)	0.524 (0.459-0.589)	0.902 (0.888-0.915)
10%	SL_ImageNet	0.816 (0.797-0.834)***	0.937 (0.925-0.949)***	0.414 (0.365-0.464)***	0.871 (0.856-0.886)***
	DINOv2	0.824 (0.807-0.841)***	0.939 (0.926-0.953)***	0.448 (0.389-0.506)***	0.871 (0.856-0.886)***
	SwAVDerm	0.811 (0.794-0.827)***	0.923 (0.907-0.938)***	0.365 (0.329-0.402)***	0.854 (0.840-0.869)***
	PanDerm	0.872 (0.855-0.888)	0.969 (0.961-0.978)	0.618 (0.550-0.686)	0.923 (0.910-0.936)
20%	SL_ImageNet	0.838 (0.820-0.855)***	0.940 (0.928-0.952)***	0.506 (0.439-0.573)***	0.880 (0.865-0.896)***
	DINOv2	0.839 (0.822-0.856)***	0.947 (0.935-0.959)***	0.534 (0.467-0.600)**	0.884 (0.869-0.899)***
	SwAVDerm	0.825 (0.809-0.842)***	0.931 (0.916-0.946)***	0.408 (0.357-0.460)***	0.871 (0.855-0.886)***
	PanDerm	0.889 (0.873-0.905)	0.975 (0.968-0.982)	0.665 (0.594-0.736)	0.935 (0.924-0.947)
30%	SL_ImageNet	0.837 (0.820-0.855)***	0.945 (0.933-0.956)***	0.505 (0.438-0.571)***	0.883 (0.867-0.899)***
	DINOv2	0.848 (0.831-0.864)***	0.951 (0.939-0.962)***	0.559 (0.488-0.630)**	0.894 (0.880-0.909)***
	SwAVDerm	0.839 (0.822-0.857)***	0.939 (0.925-0.953)***	0.493 (0.428-0.558)***	0.879 (0.864-0.895)***
	PanDerm	0.895 (0.880-0.911)	0.979 (0.973-0.985)	0.721 (0.653-0.788)	0.943 (0.931-0.954)
50%	SL_ImageNet	0.855 (0.838-0.872)***	0.957 (0.947-0.967)***	0.565 (0.492-0.638)***	0.904 (0.889-0.919)***
	DINOv2	0.855 (0.838-0.872)***	0.953 (0.940-0.965)***	0.597 (0.529-0.664)**	0.902 (0.888-0.916)***
	SwAVDerm	0.854 (0.836-0.871)***	0.953 (0.941-0.964)***	0.557 (0.489-0.626)***	0.893 (0.878-0.908)***
	PanDerm	0.909 (0.894-0.924)	0.981 (0.976-0.987)	0.749 (0.685-0.814)	0.950 (0.939-0.961)
100%	SL_ImageNet	0.872 (0.856-0.888)***	0.967 (0.958-0.976)***	0.652 (0.584-0.720)***	0.919 (0.905-0.934)***
	DINOv2	0.876 (0.860-0.892)***	0.962 (0.951-0.972)***	0.686 (0.621-0.751)**	0.913 (0.898-0.928)***
	SwAVDerm	0.864 (0.847-0.881)***	0.963 (0.954-0.972)***	0.592 (0.520-0.664)***	0.904 (0.889-0.919)***
	PanDerm	0.922 (0.908-0.936)	0.988 (0.984-0.992)	0.797 (0.744-0.850)	0.959 (0.949-0.969)

Extended Data Table 15: **Label efficiency generalization performance for dermoscopic image-based skin cancer diagnosis based on HAM_clean dataset.** Metrics: W_F1 (Weighted F1), AUROC, BACC (Balanced Accuracy), AUPR (Area Under Precision-Recall Curve). The best model for each setting is bolded. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ compared to PanDerm.

Percent	Model	W_F1	AUROC	BACC	AUPR
5%	SL_ImageNet	0.543 (0.518-0.568)***	0.815 (0.799-0.831)***	0.305 (0.272-0.338)	0.567 (0.539-0.594)***
	DINOv2	0.565 (0.541-0.590)	0.840 (0.824-0.855)**	0.305 (0.278-0.332)	0.616 (0.589-0.642)
	SwAVDerm	0.541 (0.517-0.565)***	0.809 (0.793-0.825)***	0.272 (0.248-0.297)**	0.565 (0.538-0.593)***
	PanDerm	0.586 (0.563-0.609)	0.859 (0.846-0.872)	0.304 (0.278-0.330)	0.628 (0.601-0.654)
10%	SL_ImageNet	0.561 (0.536-0.586)***	0.837 (0.822-0.851)***	0.332 (0.295-0.369)***	0.599 (0.569-0.628)***
	DINOv2	0.602 (0.577-0.627)**	0.855 (0.840-0.870)***	0.373 (0.329-0.417)*	0.642 (0.614-0.671)***
	SwAVDerm	0.571 (0.546-0.596)***	0.833 (0.818-0.848)***	0.321 (0.288-0.355)***	0.590 (0.562-0.618)***
	PanDerm	0.650 (0.626-0.675)	0.890 (0.877-0.903)	0.417 (0.374-0.459)	0.704 (0.678-0.731)
20%	SL_ImageNet	0.612 (0.587-0.638)***	0.862 (0.849-0.876)***	0.392 (0.346-0.438)*	0.647 (0.619-0.675)***
	DINOv2	0.614 (0.589-0.638)***	0.872 (0.859-0.886)***	0.374 (0.330-0.418)**	0.667 (0.638-0.695)***
	SwAVDerm	0.591 (0.566-0.615)***	0.851 (0.836-0.865)***	0.357 (0.318-0.396)**	0.629 (0.601-0.656)***
	PanDerm	0.681 (0.658-0.704)	0.910 (0.899-0.921)	0.434 (0.395-0.473)	0.735 (0.708-0.761)
30%	SL_ImageNet	0.613 (0.587-0.639)***	0.866 (0.853-0.880)***	0.426 (0.373-0.478)*	0.663 (0.636-0.690)***
	DINOv2	0.648 (0.624-0.672)***	0.880 (0.867-0.894)***	0.435 (0.381-0.489)*	0.687 (0.660-0.713)***
	SwAVDerm	0.610 (0.586-0.635)***	0.861 (0.847-0.875)***	0.343 (0.315-0.372)***	0.639 (0.611-0.668)***
	PanDerm	0.703 (0.680-0.727)	0.923 (0.913-0.933)	0.509 (0.455-0.563)	0.766 (0.742-0.790)
50%	SL_ImageNet	0.649 (0.624-0.674)***	0.888 (0.875-0.901)***	0.495 (0.435-0.556)*	0.698 (0.671-0.726)***
	DINOv2	0.660 (0.636-0.684)***	0.894 (0.883-0.906)***	0.478 (0.424-0.532)*	0.720 (0.695-0.746)***
	SwAVDerm	0.628 (0.602-0.653)***	0.880 (0.867-0.893)***	0.396 (0.352-0.439)***	0.685 (0.657-0.713)***
	PanDerm	0.733 (0.709-0.757)	0.939 (0.930-0.948)	0.578 (0.521-0.636)	0.801 (0.778-0.825)
100%	SL_ImageNet	0.698 (0.673-0.723)***	0.910 (0.898-0.922)***	0.590 (0.539-0.641)	0.747 (0.722-0.773)***
	DINOv2	0.705 (0.681-0.729)***	0.923 (0.913-0.933)***	0.565 (0.506-0.624)**	0.774 (0.748-0.800)***
	SwAVDerm	0.683 (0.659-0.708)***	0.907 (0.895-0.918)***	0.479 (0.428-0.529)***	0.733 (0.707-0.758)***
	PanDerm	0.767 (0.742-0.791)	0.951 (0.944-0.959)	0.647 (0.602-0.692)	0.843 (0.822-0.864)

Extended Data Table 16: **Label efficiency generalization performance for dermoscopic image-based skin cancer diagnosis on BCN20000 dataset.** Metrics: W_F1 (Weighted F1), AUROC, BACC (Balanced Accuracy), AUPR (Area Under Precision-Recall Curve). The best model for each setting is bolded. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ compared to PanDerm.

Percent	Model	W_F1	AUROC	BACC	AUPR
5%	SL_ImageNet	0.826 (0.802-0.851)	0.681 (0.593-0.769)*	0.548 (0.502-0.594)	0.329 (0.214-0.444)
	DINOv2	0.825 (0.798-0.853)	0.706 (0.620-0.791)	0.553 (0.502-0.605)	0.331 (0.219-0.444)*
	SwAVDerm	0.839 (0.812-0.866)	0.660 (0.567-0.754)*	0.572 (0.518-0.626)	0.355 (0.231-0.480)
	PanDerm	0.842 (0.811-0.874)	0.749 (0.663-0.834)	0.594 (0.534-0.654)	0.431 (0.299-0.562)
10%	SL_ImageNet	0.858 (0.829-0.888)	0.784 (0.708-0.861)	0.622 (0.556-0.688)	0.470 (0.345-0.595)
	DINOv2	0.853 (0.822-0.884)	0.798 (0.721-0.875)	0.626 (0.557-0.694)	0.461 (0.322-0.600)
	SwAVDerm	0.858 (0.826-0.889)	0.732 (0.647-0.816)	0.653 (0.581-0.725)	0.390 (0.263-0.517)
	PanDerm	0.857 (0.826-0.888)	0.795 (0.712-0.878)	0.652 (0.583-0.721)	0.480 (0.347-0.613)
20%	SL_ImageNet	0.847 (0.818-0.876)	0.802 (0.739-0.865)***	0.612 (0.545-0.679)*	0.458 (0.332-0.584)
	DINOv2	0.908 (0.879-0.937)	0.799 (0.726-0.833)	0.779 (0.708-0.851)	0.629 (0.500-0.758)
	SwAVDerm	0.841 (0.811-0.872)*	0.798 (0.724-0.872)**	0.607 (0.541-0.673)*	0.461 (0.333-0.589)
	PanDerm	0.879 (0.848-0.910)	0.885 (0.839-0.930)	0.718 (0.644-0.791)	0.573 (0.443-0.703)
30%	SL_ImageNet	0.851 (0.821-0.881)	0.851 (0.794-0.908)*	0.631 (0.561-0.701)*	0.506 (0.384-0.628)
	DINOv2	0.869 (0.837-0.902)	0.854 (0.794-0.914)*	0.709 (0.638-0.781)	0.580 (0.444-0.715)
	SwAVDerm	0.884 (0.854-0.915)	0.814 (0.745-0.883)**	0.679 (0.609-0.748)	0.508 (0.372-0.643)
	PanDerm	0.874 (0.842-0.906)	0.899 (0.858-0.939)	0.731 (0.657-0.805)	0.601 (0.465-0.737)
50%	SL_ImageNet	0.872 (0.842-0.903)	0.839 (0.778-0.899)**	0.679 (0.607-0.752)	0.524 (0.399-0.650)*
	DINOv2	0.882 (0.850-0.913)	0.831 (0.766-0.896)**	0.718 (0.644-0.793)	0.540 (0.403-0.677)*
	SwAVDerm	0.873 (0.841-0.904)	0.820 (0.757-0.883)**	0.687 (0.614-0.760)	0.515 (0.382-0.648)**
	PanDerm	0.890 (0.860-0.920)	0.915 (0.877-0.953)	0.750 (0.677-0.823)	0.688 (0.570-0.805)
100%	SL_ImageNet	0.873 (0.841-0.904)*	0.878 (0.823-0.933)**	0.681 (0.610-0.751)**	0.638 (0.524-0.752)*
	DINOv2	0.879 (0.848-0.911)*	0.877 (0.820-0.934)***	0.734 (0.659-0.808)	0.631 (0.507-0.754)*
	SwAVDerm	0.860 (0.828-0.891)***	0.848 (0.791-0.905)***	0.660 (0.590-0.731)**	0.538 (0.414-0.661)***
	PanDerm	0.912 (0.884-0.940)	0.949 (0.922-0.976)	0.774 (0.700-0.847)	0.771 (0.660-0.881)

Extended Data Table 17: **Label efficiency generalization performance for dermoscopic image-based melanoma detection on HIBA dataset.** Metrics: W_F1 (Weighted F1), AUROC, BACC (Balanced Accuracy), AUPR (Area Under Precision-Recall Curve). The best model for each setting is bolded. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ compared to PanDerm.

Percent	Model	W_F1	AUROC	BACC	AUPR
5%	SL_ImageNet	0.693 (0.622-0.763)	0.689 (0.602-0.777)	0.630 (0.563-0.696)	0.591 (0.483-0.700)
	DINOv2	0.697 (0.632-0.762)	0.713 (0.625-0.801)	0.653 (0.580-0.726)	0.654 (0.546-0.762)
	SwAVDerm	0.589 (0.533-0.645)***	0.609 (0.519-0.700)	0.530 (0.475-0.585)***	0.478 (0.376-0.580)*
	PanDerm	0.707 (0.642-0.772)	0.674 (0.587-0.762)	0.655 (0.585-0.725)	0.595 (0.495-0.695)
10%	SL_ImageNet	0.677 (0.610-0.743)	0.693 (0.607-0.779)**	0.617 (0.551-0.683)	0.552 (0.444-0.660)*
	DINOv2	0.722 (0.655-0.788)	0.771 (0.697-0.845)	0.683 (0.611-0.755)	0.673 (0.568-0.777)
	SwAVDerm	0.609 (0.544-0.674)***	0.619 (0.526-0.712)***	0.550 (0.477-0.622)***	0.464 (0.367-0.560)***
	PanDerm	0.738 (0.674-0.802)	0.785 (0.714-0.856)	0.682 (0.612-0.753)	0.666 (0.556-0.777)
20%	SL_ImageNet	0.672 (0.610-0.734)*	0.738 (0.655-0.821)	0.608 (0.544-0.672)*	0.601 (0.487-0.716)
	DINOv2	0.747 (0.685-0.809)	0.791 (0.715-0.867)	0.694 (0.622-0.765)	0.662 (0.553-0.770)
	SwAVDerm	0.634 (0.572-0.697)**	0.641 (0.556-0.725)**	0.574 (0.508-0.640)**	0.489 (0.392-0.586)***
	PanDerm	0.731 (0.670-0.793)	0.765 (0.691-0.838)	0.672 (0.603-0.740)	0.668 (0.565-0.771)
30%	SL_ImageNet	0.714 (0.653-0.775)	0.772 (0.696-0.849)	0.654 (0.591-0.717)	0.662 (0.544-0.781)
	DINOv2	0.728 (0.663-0.794)	0.813 (0.746-0.879)	0.681 (0.609-0.753)	0.721 (0.629-0.813)
	SwAVDerm	0.636 (0.571-0.700)**	0.644 (0.555-0.733)***	0.577 (0.507-0.647)**	0.496 (0.390-0.602)***
	PanDerm	0.729 (0.666-0.792)	0.816 (0.751-0.882)	0.676 (0.604-0.748)	0.722 (0.628-0.817)
50%	SL_ImageNet	0.720 (0.655-0.785)	0.772 (0.703-0.842)	0.661 (0.593-0.730)	0.677 (0.574-0.781)
	DINOv2	0.726 (0.661-0.792)	0.810 (0.743-0.877)	0.675 (0.602-0.747)	0.722 (0.630-0.813)
	SwAVDerm	0.686 (0.617-0.755)	0.693 (0.610-0.775)**	0.635 (0.562-0.708)	0.541 (0.436-0.647)**
	PanDerm	0.741 (0.676-0.806)	0.804 (0.739-0.870)	0.687 (0.616-0.758)	0.716 (0.625-0.808)
100%	SL_ImageNet	0.754 (0.687-0.822)	0.810 (0.744-0.875)*	0.703 (0.632-0.774)	0.704 (0.595-0.813)*
	DINOv2	0.764 (0.699-0.829)	0.797 (0.723-0.870)**	0.723 (0.653-0.793)	0.733 (0.646-0.820)
	SwAVDerm	0.747 (0.686-0.809)	0.777 (0.702-0.853)**	0.696 (0.625-0.767)	0.660 (0.544-0.776)**
	PanDerm	0.802 (0.743-0.860)	0.878 (0.824-0.931)	0.767 (0.700-0.833)	0.799 (0.714-0.884)

Extended Data Table 18: **Label efficiency generalization performance for clinical image-based melanoma detection on DermC dataset.** Metrics: W_F1 (Weighted F1), AUROC, BACC (Balanced Accuracy), AUPR (Area Under Precision-Recall Curve). The best model for each setting is bolded. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ compared to PanDerm.

Percent	Model	W_F1	AUROC	BACC	AUPR
5%	SL_ImageNet	0.566 (0.526-0.605)	0.803 (0.775-0.832)	0.395 (0.338-0.452)	0.598 (0.555-0.641)***
	DINOv2	0.560 (0.521-0.599)*	0.804 (0.776-0.833)	0.393 (0.354-0.432)***	0.607 (0.563-0.651)***
	SwAVDerm	0.468 (0.425-0.510)***	0.695 (0.661-0.729)***	0.328 (0.279-0.378)***	0.485 (0.442-0.528)***
	PanDerm	0.607 (0.569-0.646)	0.835 (0.808-0.861)	0.504 (0.432-0.576)	0.675 (0.636-0.715)
10%	SL_ImageNet	0.580 (0.540-0.621)***	0.824 (0.795-0.852)***	0.431 (0.375-0.487)**	0.630 (0.584-0.675)***
	DINOv2	0.598 (0.557-0.638)***	0.835 (0.810-0.861)***	0.432 (0.388-0.477)***	0.652 (0.608-0.695)***
	SwAVDerm	0.515 (0.474-0.556)***	0.753 (0.721-0.785)***	0.375 (0.317-0.433)***	0.550 (0.503-0.597)***
	PanDerm	0.673 (0.632-0.713)	0.877 (0.854-0.900)	0.560 (0.490-0.630)	0.733 (0.691-0.774)
20%	SL_ImageNet	0.601 (0.561-0.640)***	0.831 (0.803-0.858)***	0.486 (0.415-0.556)*	0.661 (0.616-0.705)***
	DINOv2	0.621 (0.581-0.661)**	0.843 (0.817-0.869)***	0.472 (0.408-0.536)***	0.678 (0.634-0.723)***
	SwAVDerm	0.535 (0.494-0.576)***	0.773 (0.742-0.805)***	0.398 (0.341-0.456)***	0.579 (0.535-0.623)***
	PanDerm	0.685 (0.645-0.726)	0.882 (0.860-0.904)	0.598 (0.526-0.670)	0.740 (0.699-0.781)
30%	SL_ImageNet	0.639 (0.597-0.681)***	0.854 (0.827-0.880)***	0.493 (0.426-0.560)***	0.702 (0.657-0.747)***
	DINOv2	0.636 (0.596-0.675)***	0.855 (0.830-0.880)***	0.477 (0.412-0.543)***	0.699 (0.658-0.740)***
	SwAVDerm	0.577 (0.532-0.621)***	0.806 (0.778-0.835)***	0.461 (0.390-0.532)***	0.624 (0.578-0.671)***
	PanDerm	0.729 (0.689-0.768)	0.910 (0.891-0.929)	0.645 (0.573-0.716)	0.791 (0.753-0.829)
50%	SL_ImageNet	0.634 (0.592-0.675)***	0.861 (0.835-0.887)***	0.544 (0.472-0.615)***	0.707 (0.663-0.752)***
	DINOv2	0.656 (0.614-0.698)***	0.867 (0.840-0.893)***	0.565 (0.494-0.636)***	0.716 (0.675-0.756)***
	SwAVDerm	0.595 (0.554-0.637)***	0.826 (0.798-0.854)***	0.490 (0.422-0.558)***	0.651 (0.607-0.695)***
	PanDerm	0.768 (0.730-0.806)	0.920 (0.900-0.939)	0.729 (0.665-0.793)	0.816 (0.777-0.854)
100%	SL_ImageNet	0.658 (0.617-0.699)***	0.878 (0.854-0.902)***	0.576 (0.504-0.649)***	0.747 (0.707-0.788)***
	DINOv2	0.695 (0.653-0.737)**	0.880 (0.855-0.905)***	0.575 (0.505-0.646)***	0.747 (0.706-0.787)***
	SwAVDerm	0.664 (0.624-0.704)***	0.858 (0.833-0.882)***	0.539 (0.473-0.606)***	0.709 (0.667-0.750)***
	PanDerm	0.758 (0.720-0.796)	0.931 (0.914-0.947)	0.710 (0.644-0.776)	0.844 (0.812-0.876)

Extended Data Table 19: **Label efficiency generalization performance for dermoscopic image-based skin condition classification on PAD dataset.** Metrics: W_F1 (Weighted F1), AUROC, BACC (Balanced Accuracy), AUPR (Area Under Precision-Recall Curve). The best model for each setting is bolded. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ compared to PanDerm.

Percent	Model	W_F1	AUROC	BACC	AUPR
5%	SL_ImageNet	0.805 (0.799-0.812)***	0.984 (0.983-0.985)***	0.738 (0.711-0.765)***	0.886 (0.880-0.891)***
	DINOv2	0.835 (0.829-0.841)***	0.986 (0.985-0.987)***	0.756 (0.736-0.776)***	0.905 (0.900-0.909)***
	SwAVDerm	0.738 (0.731-0.745)***	0.968 (0.966-0.969)***	0.673 (0.647-0.698)***	0.806 (0.799-0.813)***
	PanDerm	0.867 (0.861-0.872)	0.991 (0.991-0.992)	0.844 (0.836-0.853)	0.931 (0.927-0.935)
10%	SL_ImageNet	0.831 (0.824-0.837)***	0.988 (0.987-0.988)***	0.789 (0.768-0.809)***	0.908 (0.903-0.913)***
	DINOv2	0.854 (0.848-0.860)***	0.989 (0.988-0.989)***	0.790 (0.764-0.816)***	0.916 (0.911-0.921)***
	SwAVDerm	0.773 (0.766-0.780)***	0.975 (0.974-0.976)***	0.704 (0.679-0.729)***	0.842 (0.835-0.849)***
	PanDerm	0.877 (0.872-0.882)	0.992 (0.992-0.993)	0.859 (0.850-0.867)	0.937 (0.933-0.941)
20%	SL_ImageNet	0.851 (0.845-0.857)***	0.990 (0.989-0.991)***	0.799 (0.772-0.826)***	0.924 (0.919-0.928)***
	DINOv2	0.869 (0.863-0.874)***	0.991 (0.990-0.991)***	0.812 (0.787-0.838)***	0.930 (0.926-0.935)***
	SwAVDerm	0.780 (0.774-0.787)***	0.977 (0.976-0.979)***	0.708 (0.687-0.729)***	0.854 (0.848-0.861)***
	PanDerm	0.884 (0.879-0.889)	0.993 (0.993-0.994)	0.874 (0.866-0.882)	0.945 (0.942-0.949)
30%	SL_ImageNet	0.862 (0.857-0.868)***	0.991 (0.990-0.991)***	0.803 (0.778-0.827)***	0.929 (0.925-0.934)***
	DINOv2	0.879 (0.873-0.884)***	0.992 (0.991-0.992)***	0.827 (0.801-0.852)*	0.937 (0.933-0.941)***
	SwAVDerm	0.793 (0.786-0.799)***	0.980 (0.978-0.981)***	0.710 (0.702-0.719)***	0.863 (0.857-0.870)***
	PanDerm	0.896 (0.891-0.901)	0.994 (0.993-0.994)	0.866 (0.841-0.891)	0.947 (0.943-0.951)
50%	SL_ImageNet	0.865 (0.859-0.871)***	0.991 (0.990-0.992)***	0.809 (0.784-0.833)***	0.931 (0.927-0.935)***
	DINOv2	0.885 (0.879-0.890)***	0.992 (0.991-0.993)***	0.821 (0.801-0.841)***	0.938 (0.934-0.942)***
	SwAVDerm	0.798 (0.792-0.805)***	0.981 (0.980-0.982)***	0.718 (0.709-0.727)***	0.871 (0.865-0.877)***
	PanDerm	0.900 (0.895-0.904)	0.994 (0.993-0.995)	0.872 (0.848-0.897)	0.948 (0.944-0.951)
100%	SL_ImageNet	0.869 (0.863-0.875)***	0.992 (0.991-0.992)***	0.829 (0.803-0.856)***	0.934 (0.930-0.938)***
	DINOv2	0.890 (0.885-0.895)***	0.993 (0.992-0.993)***	0.820 (0.813-0.828)***	0.942 (0.938-0.947)**
	SwAVDerm	—	—	—	—
	PanDerm	0.901 (0.896-0.905)	0.994 (0.993-0.994)	0.878 (0.854-0.902)	0.947 (0.943-0.950)

Extended Data Table 20: **Label efficiency generalization performance for fine-grained skin tumor classification on PATCH16 dataset.** Metrics: W_F1 (Weighted F1), AUROC, BACC (Balanced Accuracy), AUPR (Area Under Precision-Recall Curve). The best model for each setting is bolded. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ compared to PanDerm. SwAVDerm results for the 100% setting are missing.

Dataset	Skin Tone	W_F1	Sensitivity
F17K	FST I-II (n=1195)	0.825 (0.825-0.825)	0.835 (0.835-0.835)
	FST III-IV (n=786)	0.840 (0.840-0.840)	0.851 (0.851-0.851)
	FST V-VI (n=238)	0.864 (0.864-0.864)	0.878 (0.878-0.878)
DDI	FST I-II (n=40)	0.780 (0.780-0.780)	0.750 (0.750-0.750)
	FST III-IV (n=59)	0.818 (0.818-0.818)	0.814 (0.814-0.814)
	FST V-VI (n=38)	0.854 (0.854-0.854)	0.842 (0.842-0.842)

Extended Data Table 21: **PanDerm performance across different skin tones on Fitzpatrick17k and DDI datasets.** The table shows the performance metrics for PanDerm on F17K and DDI datasets, stratified by Fitzpatrick Skin Type (FST) groups. Metrics include Weighted F1 score (W_F1) and Sensitivity. 95% CI in parentheses (identical to point estimate due to single measurement).

Category	Subgroup	n	W_F1	Sensitivity
Overall	All	1232	0.957	0.959
Sex	Female	563	0.963	0.965
	Male	659	0.951	0.953
Location	Face	70	0.862	0.871
	Lower extremity	240	0.968	0.971
	Abdomen	126	0.984	0.984
	Upper extremity	120	0.967	0.967
	Back	256	0.919	0.922
	Trunk	259	0.993	0.992
	Scalp	8	1.000	1.000
	Hand	18	1.000	1.000
	Unknown	40	1.000	1.000
	Chest	33	0.879	0.879
	Neck	18	1.000	1.000
	Foot	30	0.950	0.967
	Genital	8	1.000	1.000
Age	Old	517	0.923	0.927
	Medium	669	0.984	0.984
	Young	34	0.913	0.941

Extended Data Table 22: **Model robustness analysis across subgroups on HAM10000 dataset.** The table shows performance metrics across different subgroups based on sex, location, and age. Metrics include sample size (n), Weighted F1 score (W_F1), and Sensitivity. All metric values are point estimates.

Condition	Accuracy
Without AI assistance	0.69 (0.65-0.73)
With AI assistance	0.80 (0.76-0.84) ***

Extended Data Table 23: **Human AI collaboration performance comparison on skin cancer classification using HAM10000 dataset.** Comparison of accuracy with and without AI assistance. 95% CI in parentheses.

*** $p = 6.53 \times 10^{-8}$ compared to performance without AI assistance.

Class	Without AI	With AI	p-value	Corrected p-value
AKIEC	0.51 (0.44-0.59)	0.67 (0.60-0.74) **	0.0036	0.0254
BCC	0.76 (0.69-0.82)	0.84 (0.79-0.90)	0.0545	0.3812
BKL	0.57 (0.50-0.65)	0.77 (0.71-0.84) ***	0.0001	0.0007
DF	0.70 (0.63-0.77)	0.81 (0.75-0.87) *	0.0151	0.1059
MEL	0.69 (0.64-0.74)	0.83 (0.79-0.87) ***	<0.0001	0.0003
NV	0.83 (0.79-0.86)	0.86 (0.82-0.89)	0.1948	1.0000
VASC	0.93 (0.90-0.97)	0.95 (0.92-0.98)	0.4799	1.0000

Extended Data Table 24: **Class-specific performance comparison on human-AI collaboration study.** Comparison of accuracy with and without AI assistance for each class. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ compared to performance without AI assistance.

Experience Group	Without AI	With AI	p-value	Corrected p-value
Low	0.64 (0.58-0.70)	0.83 (0.78-0.88)**	0.0082	0.0246
Medium	0.67 (0.62-0.72)	0.79 (0.75-0.83)***	<0.0001	<0.0001
High	0.78 (0.75-0.81)	0.84 (0.82-0.86)**	0.0390	0.1170

Extended Data Table 25: **Performance by experience level on human-AI collaboration study.** Comparison of accuracy with and without AI assistance for each experience group. 95% CI in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (Bonferroni corrected) compared to performance without AI assistance.

Model	Metric	3-year	5-year	7-year
PanDerm_multi	AUC p-value	0.9463 (0.9004-0.9922) 0.0000	0.9462 (0.9125-0.9799) 0.0000	0.9303 (0.8953-0.9654) 0.0000
Multi-clinical variables	AUC p-value	0.8819 (0.7960-0.9279) 0.1601	0.9018 (0.8016-0.9420) 0.1521	0.8587 (0.7997-0.9177) 0.1475
Single clinical variables	AUC p-value	0.8396 (0.7999-0.8793) 0.0788	0.8473 (0.7859-0.9086) 0.0225	0.8021 (0.7235-0.9007) 0.0480
PanDerm_single	AUC p-value	0.9501 (0.9095-0.9908) 0.4127	0.9312 (0.8868-0.9755) 0.0871	0.9087 (0.8803-0.9371) 0.0756

Extended Data Table 26: **Survival analysis performance comparison between PanDerm and clinical variables on CombinMel dataset.** Methods include PanDerm_multi, Multi-clinical variables, Single clinical variables, and PanDerm_single. Metrics: Area Under the Time-dependent ROC Curve (AUC) at 3, 5, and 7 years with corresponding p-values. The best-performing model for each time point is bolded. 95% CI in parentheses.

Model	3-year AUC	5-year AUC	7-year AUC
PanDerm	0.9501 (0.9095-0.9908) p = 0.0000	0.9312 (0.8868-0.9755) p = 0.0000	0.9087 (0.8803-0.9371) p = 0.0000
DINOV2	0.9276 (0.8903-0.9650) p = 0.2901	0.9014 (0.8273-0.9754) p = 0.3659	0.8836 (0.7981-0.9691) p = 0.4613
SwavDerm	0.9234 (0.8715-0.9754) p = 0.2932	0.8824 (0.7943-0.9704) p = 0.2066	0.8664 (0.7776-0.9552) p = 0.2432
SL.ImageNet	0.9195 (0.8502-0.9889) p = 0.3215	0.8858 (0.7977-0.9738) p = 0.2370	0.8785 (0.7824-0.9747) p = 0.4280

Extended Data Table 27: **Survival analysis performance comparison of different models on CombinMel dataset.** Models include PanDerm, DINOV2, SwavDerm, and SL.ImageNet. Metrics: Area Under the Time-dependent ROC Curve (AUC) at 3, 5, and 7 years with corresponding p-values. The best-performing model for each time point is bolded. 95% CI in parentheses.

Model	ISIC19b	HAM_c	BCN20000	PAD	Derm7pt_c	Dermnet	MCSI	TBP_solar	Average
CLIP_base (teacher)	88.07	96.13	91.09	91.03	77.10	89.92	99.87	94.29	90.94
CLIP_large (teacher)	89.44	96.34	93.12	92.58	80.18	90.88	99.85	95.31	92.21
MONET_large (teacher)	89.05	96.65	94.03	92.55	75.17	91.41	99.85	95.59	91.79
BIOMED_CLIP_base (teacher)	74.71	82.59	80.82	88.16	60.95	85.60	97.12	80.29	81.28
Base model	86.69	96.36	92.75	94.20	82.91	93.33	99.88	95.15	92.66
+ BiomedCLIP_base	86.63	97.06	93.27	91.59	82.09	90.96	98.59	95.29	91.94
+ CLIP_base	88.78	97.21	93.55	92.42	79.50	93.13	99.87	95.63	92.51
+ CLIP_large	89.26	98.03	95.60	94.61	87.85	94.97	99.88	96.80	94.63

Extended Data Table 28: **Ablation on target representation (teacher models) across various dermatology datasets.** Models include CLIP variants, MONET, BIOMED_CLIP, and PanDerm with different pretraining strategies. Metrics represent accuracy percentages. The best-performing model for each dataset is bolded and highlighted. Datasets vary in modality and size: ISIC19b, HAM_c, BCN20000 (derm, 20k/10k/12k), PAD, Derm7pt_c, Dermnet, MCSI (clinic, 2k/839/19k/400), TBP_solar (TBP, 6k). It shows that CLIP-large pretrained on the natural domain can outperform biomedical-specific CLIP (BiomedCLIP) and dermatology-specific CLIP (MONET). This can be attributed to the limited data scale of skin images in medical domain CLIP models. Thus, CLIP-large remains the best teacher model for creating target representations for masked image modeling in dermatology. When incorporating CLIP-large as the teacher model, it significantly improved the base model (+ 1.97 on average) and also outperformed the teacher model itself (+2.42 on average).

Model	HAM_c	BCN20000	PAD	Derm7pt_c	Dermnet	MCSI	TBP_solar	Average	Training time
PanDerm (FT)	98.03	97.65	93.59	86.68	95.21	98.10	96.38	95.09	~ 80 min
PanDerm (LP)	97.40	95.19	94.50	84.94	94.36	99.53	96.09	94.57	~ 5 min
Performance difference	-0.63	-2.46	+0.91	-1.74	-0.85	+1.43	-0.29	-0.52	-75 min
Modalities	derm	derm	clinic	clinic	clinic	clinic	TBP		
Size	10k	12k	2k	839	19k	400	6k		
#class	7	9	6	2	23	4	3		

Extended Data Table 29: **Performance comparison of PanDerm (FT) and PanDerm (LP) models across various dermatology datasets.** FT: Fine-Tuning, LP: Linear Probing. Metrics represent accuracy percentages. The best-performing model for each dataset is bolded. The performance difference row shows the change from FT to LP, with positive values indicating LP outperformed FT. Datasets vary in modality, size, and number of classes as shown in the bottom rows. It shows that PanDerm using simple linear probing can perform comparably with expensive full-parameter finetuning. This suggests that PanDerm’s features are already well-suited for diverse downstream multimodal skin-related tasks without requiring further training. All models are trained and evaluated using $4 \times$ NVIDIA RTX 6000Ada GPUs.

Model	Accuracy (ACC)
PanDerm	0.932
BioMedGPT	0.866
MedViT	0.723
BiomedCLIP	0.719

Extended Data Table 30: **Model performance comparison with GMAI on skin cancer classification using HAM10000 dataset.** Accuracy (ACC) is reported for different models, including the generalist medical model BioMedGPT. PanDerm achieves the highest accuracy. Performance of baseline models are as reported in ⁶⁰.

Hyper-parameter	Value
Teacher model	CLIP
First input size	224
Second input size	196
Second interpolation	bicubic
Number of output dimensions	768
Crop min size	0.4
Crop max size	1
Patch size	16
Vocabulary size	8000
Batch size	480
Learning rate	1.5e-3
Warmup epochs	20
Total epochs	500
Gradient clipping max norm	3.0
Layer scale init value	1e-5
Color jitter	0.4
Drop path	0.2
Mask generator	block
Number of mask patches	118
Decoder layer scale init value	1e-5
Regressor depth	4
Decoder depth	0
Decoder embed dimension	1024
Decoder number of heads	16
Align loss weight	0
Latent alignment loss weight	1
Number of GPUs	4
Distributed launch	torchrun
Processes per node	4

Extended Data Table 31: **PanDerm hyperparameters used in pretraining.** $4 \times 80\text{GB}$ NVIDIA H100 GPUs were used for pretraining.

Hyperparameter	Value
Batch size	256
Epochs	50
learning rate	5e-4
Layer decay	0.75
Weight decay	0.05
Drop path	0.2
Reprob	0.25
Mixup	0.8
Cutmix	1.0

Extended Data Table 32: **PanDerm hyperparameters used in finetuning.** Single 49GB NVIDIA 6000Ada GPU was used for downstream finetuning.

References

1. Kittler, H., Pehamberger, H., Wolff, K. & Binder, M. Diagnostic accuracy of dermoscopy. *The Lancet. Oncology* **3** 3, 159–65 (2002). URL <https://api.semanticscholar.org/CorpusID:25669479>.
2. Primiero, C. A. *et al.* A narrative review: opportunities and challenges in artificial intelligence skin image analyses using total body photography. *Journal of Investigative Dermatology* (2024).
3. Primiero, C. A. *et al.* A protocol for annotation of total body photography for machine learning to analyze skin phenotype and lesion classification. *Frontiers in Medicine* **11**, 1380984 (2024).
4. Olsen, C. M. *et al.* Risk stratification for melanoma: models derived and validated in a purpose-designed prospective cohort. *JNCI: Journal of the National Cancer Institute* **110**, 1075–1083 (2018).
5. Usher-Smith, J. A., Emery, J., Kassianos, A. P. & Walter, F. M. Risk prediction models for melanoma: a systematic review. *Cancer epidemiology, biomarkers & prevention* **23**, 1450–1463 (2014).
6. Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. *Nature medicine* **26**, 900–908 (2020).
7. Yap, J., Yolland, W. & Tschandl, P. Multimodal skin lesion classification using deep learning. *Experimental dermatology* **27**, 1261–1267 (2018).
8. Luo, N. *et al.* Artificial intelligence-assisted dermatology diagnosis: from unimodal to multimodal. *Computers in Biology and Medicine* 107413 (2023).
9. Rapini, R. P. *Practical dermatopathology* (Elsevier Health Sciences, 2021).
10. Song, A. H. *et al.* Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering* **1**, 930–949 (2023).
11. Kittler, H. *et al.* Identification of clinically featureless incipient melanoma using sequential dermoscopy imaging. *Archives of dermatology* **142**, 1113–1119 (2006).
12. Altamura, D., Avramidis, M. & Menzies, S. W. Assessment of the optimal interval for and sensitivity of short-term sequential digital dermoscopy monitoring for the diagnosis of melanoma. *Archives of dermatology* **144**, 502–506 (2008).
13. Todorovic, D. *et al.* Dermatoscopic patterns of cutaneous metastases: A multicentre cross-sectional study of the international dermoscopy society. *Journal of the European Academy of Dermatology and Venereology* **38**, 1432 – 1438 (2024). URL <https://api.semanticscholar.org/CorpusID:268382698>.
14. Lallas, K. *et al.* Prediction of melanoma metastasis using dermatoscopy deep features: An international multicenter cohort study. *Journal of Clinical Oncology* (2024). URL <https://api.semanticscholar.org/CorpusID:270989767>.
15. Gui, H., Omiye, J. A., Chang, C. T. & Daneshjou, R. The promises and perils of foundation models in dermatology. *Journal of Investigative Dermatology* (2024).
16. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
17. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering* **6**, 1346–1352 (2022).
18. Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D. & Makedon, F. A survey on contrastive self-supervised learning. *Technologies* **9**, 2 (2020).

19. He, K. *et al.* Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009 (2022).
20. Zhou, Y. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
21. Pai, S. *et al.* Foundation model for cancer imaging biomarkers. *Nature machine intelligence* **6**, 354–367 (2024).
22. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* 1–8 (2024).
23. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**, 850–862 (2024).
24. Vorontsov, E. *et al.* A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine* (2024). URL <https://api.semanticscholar.org/CorpusID:271332336>.
25. Wang, X. *et al.* A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* 1–9 (2024).
26. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical ai. *Nature Medicine* **28**, 1773–1784 (2022).
27. Azizi, S. *et al.* Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering* **7**, 756–779 (2023).
28. Azizi, S. *et al.* Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3478–3488 (2021).
29. Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
30. Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12104–12113 (2022).
31. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
32. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**, 211–252 (2015).
33. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, 843–852 (2017).
34. Wang, Z., Lyu, J. & Tang, X. Autosmim: Automatic superpixel-based masked image modeling for skin lesion segmentation. *IEEE Transactions on Medical Imaging* (2023).
35. Tschandl, P., Rosendahl, C. & Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5**, 1–9 (2018).
36. Shen, Y. *et al.* Optimizing skin disease diagnosis: harnessing online community data with contrastive learning and clustering techniques. *NPJ Digital Medicine* **7**, 28 (2024).
37. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).
38. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

39. Hou, Z., Sun, F., Chen, Y.-K., Xie, Y. & Kung, S.-Y. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049* (2022).
40. Oquab, M. *et al.* DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024). URL <https://openreview.net/forum?id=a68SUt6zFt>.
41. Kim, C. *et al.* Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine* 1–12 (2024).
42. Zhang, S. *et al.* Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* (2023).
43. Pacheco, A. G. *et al.* Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief* **32**, 106221 (2020).
44. Yan, S. *et al.* Epvt: Environment-aware prompt vision transformer for domain generalization in skin lesion recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 249–259 (Springer, 2023).
45. Yan, S. *et al.* Prompt-driven latent domain generalization for medical image classification. *IEEE Transactions on Medical Imaging* (2024).
46. Dermnet. Dermnet (2023). <https://dermnet.com/>.
47. Yu, Z. *et al.* Early melanoma diagnosis with sequential dermoscopic images. *IEEE Transactions on Medical Imaging* **41**, 633–646 (2021).
48. Zhang, B. *et al.* Short-term lesion change detection for melanoma screening with novel siamese neural network. *IEEE transactions on medical imaging* **40**, 840–851 (2020).
49. Sacchetto, L. *et al.* Skin melanoma deaths within 1 or 3 years from diagnosis in europe. *International Journal of Cancer* **148**, 2898–2905 (2021).
50. Kurtansky, N. R. *et al.* The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection. *Scientific Data* **11**, 884 (2024).
51. Arnold, M. *et al.* Global burden of cutaneous melanoma attributable to ultraviolet radiation in 2012. *International journal of cancer* **143**, 1305–1314 (2018).
52. Primiero, C. A. *et al.* Evaluation of the efficacy of 3d total-body photography with sequential digital dermoscopy in a high-risk melanoma cohort: protocol for a randomised controlled trial. *BMJ open* **9**, e032969 (2019).
53. Koh, U. *et al.* ‘mind your moles’ study: protocol of a prospective cohort study of melanocytic naevi. *BMJ open* **8**, e025857 (2018).
54. Grob, J. & Bonerandi, J. The ‘ugly duckling’ sign: identification of the common characteristics of nevi in an individual as a basis for melanoma screening. *Archives of dermatology* **134**, 103–104 (1998).
55. Milton, M. A. A. Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. *arXiv preprint arXiv:1901.10802* (2019).
56. Ma, J. *et al.* Segment anything in medical images. *Nature Communications* **15**, 654 (2024).
57. Codella, N. C. *et al.* Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, 168–172 (IEEE, 2018).

58. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
59. Tu, T. *et al.* Towards generalist biomedical ai. *NEJM AI* **1**, A10a2300138 (2024).
60. Zhang, K. *et al.* A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine* 1–13 (2024).
61. Ye, Y. *et al.* Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11114–11124 (2024).
62. Daneshjou, R. *et al.* Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances* **8**, eabq6147 (2022).
63. Groh, M. *et al.* Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nature Medicine* **30**, 573–583 (2024).
64. Groh, M. *et al.* Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1820–1828 (2021).
65. Tschandl, P. *et al.* Human–computer collaboration for skin cancer recognition. *Nature medicine* **26**, 1229–1234 (2020).
66. Winkler, J. K. *et al.* Assessment of diagnostic performance of dermatologists cooperating with a convolutional neural network in a prospective clinical study: human with machine. *JAMA dermatology* **159**, 621–627 (2023).
67. Chanda, T. *et al.* Dermatologist-like explainable ai enhances trust and confidence in diagnosing melanoma. *Nature Communications* **15**, 524 (2024).
68. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
69. Guan, J., Gupta, R. & Filipp, F. V. Cancer systems biology of tcga skcm: efficient detection of genomic drivers in melanoma. *Scientific reports* **5**, 7857 (2015).
70. Kriegsmann, K. *et al.* Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology* **12**, 1022967 (2022).
71. Zhang, X. *et al.* Cae v2: Context autoencoder with clip latent alignment. *Transactions on Machine Learning Research* (2023).
72. Peng, Z., Dong, L., Bao, H., Ye, Q. & Wei, F. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366* (2022).
73. Pewton, S. W. & Yap, M. H. Dark corner on skin lesion image dataset: Does it matter? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4831–4839 (2022).
74. Bradski, G., Kaehler, A. *et al.* Opencv. *Dr. Dobb's journal of software tools* **3** (2000).
75. Alcantarilla, P. F. & Solutions, T. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell* **34**, 1281–1298 (2011).
76. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**, 555–570 (2021).

77. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136 (PMLR, 2018).
78. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
79. Lin, X., Yu, L., Cheng, K.-T. & Yan, Z. Batformer: Towards boundary-aware lightweight transformer for efficient medical image segmentation. *IEEE Journal of Biomedical and Health Informatics* **27**, 3501–3512 (2023).
80. Groger, F. *et al.* Towards reliable dermatology evaluation benchmarks. *ArXiv* **abs/2309.06961** (2023). URL <https://api.semanticscholar.org/CorpusID:261705763>.
81. Hernández-Pérez, C. *et al.* Bcn20000: Dermoscopic lesions in the wild. *Scientific Data* **11**, 641 (2024).
82. Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics* **23**, 538–546 (2018).
83. Mendonça, T., Celebi, M., Mendonca, T. & Marques, J. Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy image analysis* **2** (2015).
84. Giotis, I. *et al.* Med-node: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications* **42**, 6578–6585 (2015).
85. Clark, K. *et al.* The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging* **26**, 1045–1057 (2013).
86. Kahler, S. *et al.* Automated photodamage assessment from 3d total body photography for an objective assessment of melanoma risk. *Australasian Journal of Dermatology* **64**, 8–8 (2023).