

# When Dermatology AI Fails Darker Skin: Foundation-Model Embeddings Recover Diagnostic Signal on the Diverse Dermatology Images Benchmark

Riddhi Pritam Singhvi

Stillwater Area High School, Oak Park Heights, Minnesota, USA

## Corresponding authors

Riddhi Pritam Singhvi, Stillwater Area High School, Oak Park Heights, Minnesota, USA.

**Received:** September 15, 2026; **Accepted:** September 22, 2026; **Published:** October 01, 2026

## Abstract

**Background:** Skin tone can affect how well dermatology AI classifiers perform. When a classifier does poorly on a subgroup, it is hard to tell whether the image itself lacks the information needed for diagnosis or whether the model fails to use the information that is there.

**Methods:** We tested four configurations on 656 biopsy-confirmed lesions in the Diverse Dermatology Images (DDI) dataset: native Deep Derm, a recalibrated Deep Derm, and logistic probes trained on frozen Derm Foundation and DINOv2 embeddings. All used near-duplicate-grouped nested five-fold cross-validation. We measured performance by subgroup AUC, predictive information, and how many cancers Deep Derm caught at a fixed threshold.

**Results:** At a threshold set for 90% specificity in light skin, native Deep Derm flagged 13 of 49 light-skin cancers but caught none of the 48 dark-skin cancers. Dark-skin AUC was near chance (0.538; 95% CI, 0.453–0.626). Rescaling the scores did not help (AUC 0.510; 95% CI, 0.422–0.604). Both Derm Foundation (AUC 0.807; 95% CI, 0.727–0.881) and DINOv2 (AUC 0.786; 95% CI, 0.706–0.863) did substantially better on dark skin, raising AUC by 0.269 and 0.248 over Deep Derm. There was not enough data to tell whether predictive information was the same across skin groups.

**Interpretation:** Frozen foundation model embeddings could distinguish cancers from benign lesions in DDI's images, while Deep Derm's released scores carried almost no useful ranking information for dark skin. Because swapping in a foundation model also swaps out the classifier, and Deep Derm's internal features were not available, the experiment cannot pin the difference on representation alone. These results matter for auditing subgroup performance, not for deploying any of these models clinically.

**Keywords:** Dermatology AI, Skin Cancer, Health Equity, Foundation Model, Self-Supervised Learning, Skin Tone

## Introduction

Melanoma is a serious public health problem, and it affects Black and White populations differently. Black patients are about 2.5 times more likely to be diagnosed at stage IV than White patients [1]. Survival rates are also lower: for 2013–2019, five-year melanoma survival was 71% for Black patients and 94% for White patients [2]. The economic cost is also high. About 6.1 million adults in the U.S. are treated for basal- and squamous-cell carcinomas each year, at an estimated cost of \$8.9 billion. These cancers go beyond what DDI covers, but they represent a large clinical pathway where image-based triage could be used. If the system fails for a subgroup, the consequences are real: missed referrals, delayed biopsies, and added costs for patients already diagnosed late.

These differences are not just biological. Later diagnoses can result from limited access to dermatologic care, low screening rates, and lesions showing up in places that are examined less often. An automated screening system could help close that gap or make it worse. So looking at overall performance is not enough. Any tool meant to expand access must be tested where inequities are greatest.

Deep learning has shown promise for classifying skin lesions, though accuracy varies across populations and imaging conditions [3]. Introduced the DDI dataset and showed that Deep Derm performed significantly worse on darker skin [4]. Found the same pattern in Fitzpatrick17k [5]. These studies showed that the problem exists. They did not explain why.

This question also matters for regulation. As AI-based lesion assessment moves toward clinical use, the FDA has noted that

**Citation:** Riddhi Pritam Singhvi. When Dermatology AI Fails Darker Skin: Foundation-Model Embeddings Recover Diagnostic Signal on the Diverse Dermatology Images Benchmark. *J Clin Dermatol Skin Care*. 2026. 1(1): 1-7. DOI: doi.org/10.61440/JCDSC.2026.v1.06

these systems need to be tested across the full U.S. population, since skin lesions can look different depending on skin tone [6]. This study does not evaluate a cleared device. It addresses a question that comes up in device review: can a system perform well overall and still fail in a clinically important subgroup? DDI is well suited to this question because it covers all Fitzpatrick groups and every diagnosis is biopsy-confirmed. Biopsy confirmation reduces one key concern: labels might be weaker for some groups, while keeping images and labels consistent across model comparisons. That said, DDI is still a curated benchmark, not a representative clinical population.

A low subgroup AUC could mean one of two things. Either the image lacks enough visual information to make a diagnosis, or the information is there, but the model is not picking it up. The fix depends on which one it is. If the image is the problem, the camera or lighting needs improvement. If the model is the problem, the model needs better auditing.

This study pulls apart those two explanations by testing four setups on 656 DDI images: Deep Derm's own scores, a rescaled version of those scores, and simple classifiers built on frozen features from Derm Foundation and DINOv2. Derm Foundation was trained on dermatology images, while DINOv2 was trained on general images with no skin cancer labels. If either can tell cancers from benign lesions in the dark-skin group, then the pictures already contain the needed information, and Deep Derm is failing to use it.

Each of these four setups answers a separate question, so they are not meant to be compared head to head. Recalibration checks whether rescaling Deep Derm's scores helps. The foundation-model probes check whether reading the image differently does better. Because the probes also use a new classifier, the experiment cannot say which change made the difference.

**Table 1: Maps Each Check to the Question it Asks and what it can tell us.**

| Stage                  | Question                                                         | Tool                                                             | Permitted inference                                                              |
|------------------------|------------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------------------------------|
| 1. Establish failure   | Does the benchmark classifier fail for dark-skinned individuals? | Subgroup AUC; fixed-threshold sensitivity                        | Quantifies the disparity and its operating-point consequence                     |
| 2. Test scalar repair  | Is the released score merely miscalibrated?                      | Cross-validated one-feature logistic probe                       | Tests whether monotonic rescaling recovers ranking                               |
| 3. Test recoverability | Do the same images contain usable signal?                        | Frozen Derm Foundation and DINOv2 embeddings                     | Shows whether an alternative representation-plus-head can recover discrimination |
| 4. Audit robustness    | Could leakage or one split explain the gain?                     | Near-duplicate grouping; nested CV; five seeds; paired bootstrap | Reduces several optimistic-bias pathways and quantifies conditional uncertainty  |
| 5. Check the loop      | Has discrimination also become reliable probability information? | AUC plus predictive information                                  | Separates ranking recovery from calibration-sensitive subgroup parity            |

## Dataset

DDI contains 656 clinical photographs from 570 patients: 208 light skin (Fitzpatrick I-II), 241 medium skin (Fitzpatrick III-IV), and 207 dark skin (Fitzpatrick V-VI) [4]. Of these, 171 (26.1%) are malignant, split as 49 light, 74 medium, and 48 dark. The dataset came from the Stanford AIMI Shared Datasets repository. Patient identifiers were not included in the release we analyzed, so we cannot guarantee that images from the same patient ended up in separate folds.

## Why Representation was Easy to Miss

Data imbalance was an obvious early concern in medical imaging fairness. If training sets have few examples of darker skin, collecting and reweighting diverse examples is a straightforward fix. But that fix alone is not enough. Once a classifier has lost the ability to rank a subgroup correctly, no threshold adjustment or score rescaling can get it back. And if you only look at the final output, it is easy to mistake that model limitation for a problem with the images themselves.

Foundation models offer a way to test this without retraining a large network. Frozen embeddings can be extracted from the same images, and a small classifier trained the same way can measure how much useful information each feature set actually holds. This approach is not meant to replace better training data, and it does not prove that any one representation will work in a clinical setting. It helps decide the next step: whether to stop at recalibration, move to richer representations, or go back and improve image collection.

## Methods Study Design

This study is a post-observational analysis of a published clinical-image dataset, examining how model performance differed across the three Fitzpatrick groups. Using a fixed benchmark, it asks whether different model configurations can find a malignancy signal in the available images. The study says nothing about how these models would perform on new patients, new devices, or new imaging conditions. The results speak to how well different representations work on DDI, not to whether any of them are ready for clinical use.

## Analytic Toolbox

The analysis runs through a series of checks, each asking a different question. The answer to one does not depend on the answer to another.

All DDI diagnoses were biopsy-confirmed, which makes the labels stronger than in datasets that rely on visual diagnosis alone. Biopsy confirmation does not remove all uncertainty, since site selection, sampling, and pathological interpretation can still vary. But it does make label quality a less likely explanation for the model differences seen here.

Images with a DINOv2 embedding cosine similarity above 0.95 were flagged as near-duplicates. That grouping produced 651 clusters from 656 images. All images in the same cluster were kept in the same cross-validation fold, so near-duplicates never appeared in both the training and test sets. We based the grouping on image similarity only because patient identifiers were not available.

Fitzpatrick type is a clinical phototype, not a direct measure of melanin or tissue optics. Following the original DDI analysis, skin types were grouped as light (I-II), medium (III-IV), and dark (V-VI). Differences across these groups could reflect pigmentation, diagnosis mix, anatomic site, imaging conditions, or some combination. The analysis cannot separate the biological effect of pigmentation from these other factors.

DDI's skin-tone balance makes subgroup estimates possible where most dermatology datasets cannot. That said, the groups are not guaranteed to have equally difficult cases. Diagnosis, lesion size, anatomic site, and photo quality may all differ across groups, and those differences can affect subgroup AUC.

## Models

Deep Derm was the main classifier tested in the original DDI study, so we used it as the benchmark here as well [4]. Two versions were tested. The native version used Deep Derm's released malignancy scores as is. The recalibrated version fed those scores into a logistic regression trained within each cross-validation fold to see whether rescaling could improve performance. Neither version looks inside Deep Derm's hidden features.

We chose Derm Foundation and DINOv2 because they were trained in very different ways. Derm Foundation was built specifically for dermatology images. DINOv2 is a general-purpose vision model with no connection to skin cancer diagnosis. If both recover signal in dark skin, the result is less likely to depend on any one architecture or any one training approach.

A general medical-imaging foundation model would have been a useful middle ground, but fell outside the scope of this study. The two models chosen give a broad comparison between dermatology-specific and general-purpose representation.

We used both models as frozen feature extractors, meaning their weights were not updated on DDI. We trained only a regularized logistic regression on top. Freezing the encoders kept the number of trainable parameters small, which makes the comparison more stable but does not eliminate overfitting. To guard against that, all reported results come from held-out predictions, and hyperparameter tuning was nested inside each training fold.

Derm Foundation produces 6144-dimensional embeddings and was built for dermatology image tasks [7]. It is designed to

learn features that carry clinical meaning across different skin conditions. Here, the question is whether those features can help classify malignancies in the skin types found in DDI.

DINOv2 Vi T-L/14 is a self-supervised vision transformer trained on LVD-142M, a collection of 142 million curated web images [8]. Because the contents of LVD-142M are not public, it is impossible to know how many medical images it saw during training. DINOv2 was never designed or trained for skin cancer classification, making it a useful test of whether general visual features transfer to this task [9].

## Evaluation Protocol

We evaluated all three probes using the same cross-validation setup. Each image received a held-out prediction through a stratified, grouped five-fold procedure, repeated across five random seeds (0-4). Within each outer training fold, a three-fold inner cross-validation selected the regularization strength for logistic regression from  $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ . We used only training data to fit the scaler and the logistic regression. We averaged each image's five predictions into a single score, then calculated AUC from those averaged scores.

All learned probes used the same splits, tuning grid, preprocessing, and evaluation code. Native Deep Derm differed because its scores came straight from the model, with no fitting on DDI. Because recalibration learned a slightly different mapping in each fold, pooling predictions across folds can shift rankings slightly. All recalibration coefficients were positive.

When a foundation model is swapped in, both the image features and the classifier change at the same time. Deep Derm's original classifier is replaced by a logistic probe trained on a high-dimensional embedding. The one-dimensional Deep Derm probe retrains just the classifier on DDI while keeping the original score as a partial control. A cleaner test would use Deep Derm's internal features with the same logistic probe, but those features were not available.

## Predictive Information

The main metric for measuring disparity is cross-validated predictive information.  $PI = H(Y - CE(Y_{test}, f(S)))H(Y)$  is the binary entropy of malignancy prevalence in the outer training data. CE is the cross-entropy between held-out labels and predictions in a subgroup. Because the logistic regression pipeline already produced calibrated predictions, we needed no additional calibration. We used the same prevalence across training folds so differences in PI reflect both ranking ability and calibration. We pooled held-out predictions before computing cross-entropy for each subgroup.

$$\Delta PI = PI_{\text{dark}} - PI_{\text{light}}$$

AUC measures whether the model ranks cancers above benign lesions. PI goes further: it also penalizes a model that ranks correctly but assigns the wrong probabilities. A model can score well on AUC and still do poorly on PI if its probability estimates are off.

The main comparison is  $\Delta PI$ , the difference in predictive information between dark- and light-skin groups. A negative

$\Delta$ PI means the model carries less predictive information for dark skin. Across seeds, the standard deviation of dark-skin AUC was 0.02 for native Deep Derm and 0.03 for the foundation-model probes.

We estimated uncertainty using 2,000 bootstrap replicates of the averaged out-of-fold predictions (seed 42). These intervals capture uncertainty in the predictions themselves but not uncertainty from model fitting or analytical choices.

A confidence interval on  $\Delta$ PI that includes zero means the sample was too small to confirm a deficit. That does not mean the groups are equal. Proving equality would require a pre-specified margin and a formal equivalence test.

AUC and PI are reported together for a reason. AUC only measures ranking. PI also accounts for whether the predicted probabilities are accurate. If both agree, the model ranks well and produces reliable probabilities. If AUC is high but PI lags, calibration is the concern.

### Operating-Point Analysis

We chose a threshold from native Deep Derm's scores to achieve roughly 90% specificity in light skin, then applied it unchanged to medium and dark skin. Results are reported as sensitivity with Clopper-Pearson confidence intervals. This threshold is not a validated clinical cutoff; it shows what happens when one decision rule is applied across groups. We did not perform threshold analysis for the foundation-model probes, which we compared using AUC alone.

The 90% specificity level accepts some false positives to catch more cancers. We chose it to illustrate the consequences of a single aggregate rule, not to recommend a clinical threshold or optimize net benefit.

### Software and Reproducibility

Analyses used scikit-learn 1.3, Py Torch 2.1, and Python 3.10. Cross-validation seeds were 0-4, the bootstrap seed was 42, the near-duplicate cosine similarity threshold was 0.95 and the regularization grid was  $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ . Preprocessing, embedding extraction, model fitting and fold assignment were all done reproducibly

## Results

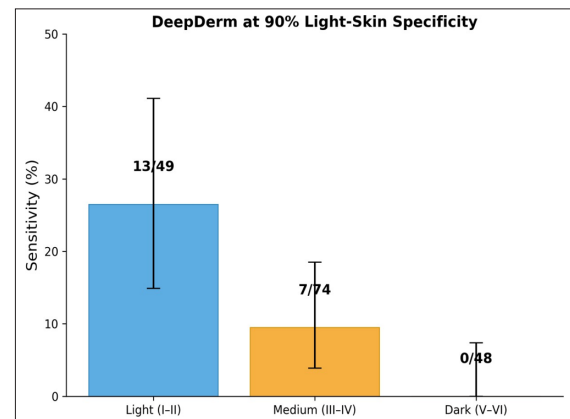
### Deep Derm: Failure on Dark Skin

Native Deep Derm's dark-skin AUC was 0.538 (95% CI, 0.453-0.626), barely better than chance. At the threshold set for roughly 90% specificity in light skin, sensitivity was 26.5% for light-skin cancers (13/49; 95% CI, 14.9%-41.1%), 9.5% for medium-skin cancers (7/74; 95% CI, 3.9%-18.5%), and 0% for dark-skin cancers (0/48; 95% CI, 0%-7.4%) (Figure 1).

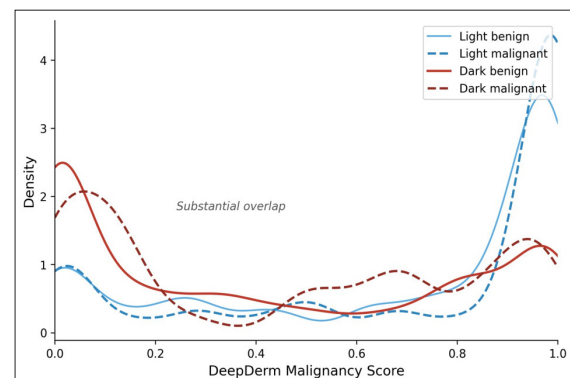
The same threshold that caught 13 light-skin cancers caught none of the 48 dark-skin cancers. The 0% sensitivity is specific to DDI and the chosen threshold, and does not mean Deep Derm can never detect dark-skin cancers. But the finding does show that a decision rule tuned for one group can completely miss another.

The score distributions for dark-skinned benign and malignant lesions overlapped heavily (Figure 2), consistent with the AUC hovering near chance.

That overlap suggests moving the threshold alone would not fix the problem. If the scores carried useful ranking information but were just miscalibrated, benign and malignant lesions would still separate. The next step tested whether rescaling the scores could improve discrimination.



**Figure 1:** Deep Derm Sensitivity at 90% Light-Skin Specificity, by Fitzpatrick Group.



**Figure 2:** Deep Derm Score Distributions by Skin Tone and Malignancy Status.

### Recalibration does not Recover Performance

Rescaling Deep Derm's scores with a cross-validated logistic probe did not improve dark-skin discrimination. Dark-skin AUC after recalibration was 0.510 (95% CI, 0.422-0.604), a change of 0.028 (95% CI, -0.062 to 0.003) relative to native Deep Derm. Rescaling could not restore the model's ability to rank dark-skin lesions.

The failure of rescaling rules out simple miscalibration, but does not pinpoint which part of Deep Derm is responsible. Deep Derm's released probability is a single number, and collapsing the model's output to one number may throw away information that still exists in the hidden layers. Without access to Deep Derm's internal features, a fair comparison with the high-dimensional foundation-model embeddings is not possible.

### Foundation-Model Representation Recovery

Both frozen foundation models, tested with the same probe setup, performed substantially better on dark skin than Deep Derm (Table 2; Figure 3). Derm Foundation reached a dark-skin

AUC of 0.807 (95% CI, 0.727-0.881), a gain of 0.269 (95% CI, 0.165-0.380). DINOv2 reached 0.786 (95% CI, 0.706-0.863), a gain of 0.248 (95% CI, 0.141-0.360). Both gains were clearly above zero.

Table 2: AUC by Skin Tone Across All Four Configurations (Near-Duplicate-Grouped Cross-Validation)

| Configuration            | Light AUC [95% CI]   | Medium AUC [95% CI]  | Dark AUC [95% CI]    | Dark gain [95% CI]      |
|--------------------------|----------------------|----------------------|----------------------|-------------------------|
| Deep Derm (native)       | 0.615 [0.515, 0.714] | 0.614 [0.536, 0.689] | 0.538 [0.453, 0.626] | -                       |
| Deep Derm (recalibrated) | 0.523 [0.431, 0.612] | 0.605 [0.529, 0.682] | 0.510 [0.422, 0.604] | -0.028 [-0.062, +0.003] |
| Derm Foundation (probe)  | 0.768 [0.684, 0.848] | 0.856 [0.801, 0.907] | 0.807 [0.727, 0.881] | +0.269 [+0.165, +0.380] |
| DINOv2 (probe)           | 0.747 [0.665, 0.823] | 0.809 [0.746, 0.867] | 0.786 [0.706, 0.863] | +0.248 [+0.141, +0.360] |

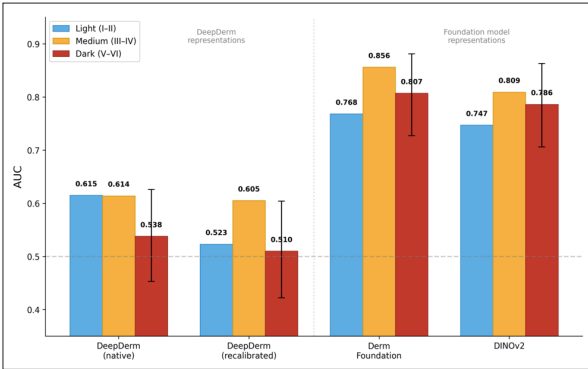


Figure 3: AUC Across all three Fitzpatrick Groups for Each Configuration.

Since the images and labels stayed the same across all setups, the near-chance dark-skin AUC is not limited by the pictures. The foundation-model embeddings pulled useful information from the same photos that Deep Derm could not use. Because swapping in a foundation model also swaps the classifier, the experiment supports a combined explanation involving both the image features and the classifier. Still, it cannot say which one mattered more.

Medium-Skin Performance

Foundation-model embeddings also improved performance in the medium-skin subgroup. Derm Foundation reached an AUC of 0.856 (95% CI, 0.801-0.907) and DINOv2 reached 0.809 (95% CI, 0.746-0.867), both well above Deep Derm's 0.614 (95% CI, 0.536-0.689). Derm Foundation's highest AUC was in the medium-skin subgroup.

Performance did not simply drop as skin got darker. The pattern suggests that AUC differences are not purely a pigmentation effect. Diagnosis mix, anatomic site, imaging conditions, and sampling variation can all affect how difficult the task is within each group.

Predictive Information

For both foundation-model probes, the dark-minus-light PI difference was small but imprecise: Derm Foundation ΔPI = -0.031 (95% CI, -0.162 to 0.092) and DINOv2 ΔPI = -0.047 (95% CI, -0.151 to 0.054). Both confidence intervals were wide and included zero.

With only 48 dark-skin cancers, the sample was too small to tell whether the groups performed equally or whether a small deficit existed. Confirming equality would require more data and a formal equivalence test.

Discussion

On DDI, frozen Derm Foundation and DINOv2 embeddings separated dark-skin cancers from benign lesions far better than native Deep Derm. Rescaling Deep Derm's released scores did not help. The problem was not a badly chosen threshold. Deep Derm's score did not carry enough ranking information for dark skin, while other combinations of image features and classifiers did, using the same images.

DINOv2 is especially telling because it was never built for skin cancer. The fact that a general purpose vision model can detect malignancy in dark-skin images suggests the information is in the pictures. Deep Derm's failure is not just about the images being hard to read.

DINOv2 is not a true negative control because its web-scale training data may include medical images. Still, it never saw DDI's malignancy labels during training, and its frozen features transferred well to the task. That likely reflects the morphology, texture, color, and shape information that modern self-supervised methods pick up.

showed that dermatology classifiers can have large performance gaps across skin tones[4]. The present study takes a different step by asking whether the model can use information already present in the images. Both frozen embedding spaces showed strong discrimination, including in the medium-skin group.

Larger, more representative dermatology datasets are still needed, especially for external and prospective testing. But representation can be tested on existing benchmarks. A quicker way to tell whether a subgroup ceiling comes from the images or the model is to test several fixed representations on the same folds before committing to a bigger intervention.

These results do not prove that Deep Derm's disparities are purely a representation problem, that Derm Foundation and DINOv2 are fair, or that any of the configurations tested are safe for clinical use. The predictive information estimates were underpowered so that they could look different with more data. The case mix across Fitzpatrick groups may also differ, and every probe was built and tested on a single curated dataset. Any claims about generalization would need independent evidence.

Deep Derm's disparities could come from its convolutional backbone, its classification head, the interaction between the two, its training data, or its optimization objective. Pinpointing the representation as the cause would require testing Deep Derm's own internal features with the same logistic probe, and those features were not available.

## Clinical and Regulatory Implications

Overall performance is not enough for safety evaluation. At 90% specificity in light skin, all 48 dark-skin cancers were missed. Developers and regulators should report performance for clinically important skin-tone subgroups with confidence intervals, test both ranking and calibration before approving, and move beyond Fitzpatrick type to better measures of skin tone.

Diagnosing algorithmic disparities takes more than reporting subgroup numbers. In the present study, the images, labels, splits, and evaluation stayed the same while the model configuration changed. That does not prove causation, but it does rule out a lack of information in the images as the explanation.

AUC and PI answer different questions. AUC measures ranking. PI also accounts for whether predicted probabilities are accurate. The foundation models showed clear AUC gains, while the PI comparisons were too imprecise to draw firm conclusions with only 48 dark-skin cancers. The gap between AUC and PI reflects the small sample size, not a contradiction.

## Strengths and Scope

Three design choices strengthen the findings. First, we tested all four configurations on the same biopsy-confirmed images, so differences in results cannot come from differences in test data. Second, all learned probes shared the same folds, hyperparameter tuning, scaling, and random seeds, so bootstrap comparisons match predictions image by image. Third, we used two very different foundation models, so the recovery does not depend on a single dermatology-specific system. Their agreement points to recoverable signal in the images, though it does not identify which features are responsible.

The study is deliberately narrower than a clinical validation. It asks two questions: is Deep Derm's dark-skin failure built into the photographs, and can rescaling fix it? The answer to both is no. The study does not say which system should be deployed, what threshold to use for biopsy, or whether these gains would hold in a different setting. Those questions need new datasets and separate evaluation.

## Limitations

- No patient identifiers: The DDI release we used did not include patient identifiers. Near-duplicate grouping reduced image-level leakage, but different lesions or views from the same patient may still appear in different folds.
- Conditional uncertainty: Bootstrap intervals are based on a fixed set of averaged out-of-fold predictions. They do not account for variation from fold creation, model refitting, or other analytical choices so they may understate the true uncertainty.
- Small dark-skin cancer subgroup: Only 48 dark-skin cancers were available. That limits intergroup comparisons, especially for sensitivity and PI, and the point estimates should not be treated as definitive.
- Curated dataset: DDI was assembled for research and may not capture the variety seen in everyday clinical practice, such as different cameras, lighting, backgrounds, disease rates, and referral reasons.
- Unknown DINOv2 pretraining composition: LVD-142M is

not fully available to the public. Because the contents of LVD-142M are not fully public, it is impossible to know how many medical images it included, where they came from, or whether consent was obtained.

- Case-mix confounding: The skin-tone groups may differ in diagnosis, anatomic site, and imaging conditions, and the data do not allow these factors to be separated from the effect of pigmentation.

## Conclusion

At a threshold tuned for light skin, Deep Derm missed all 48 dark-skin cancers. Rescaling the scores did not help. Derm Foundation and DINOv2 raised the dark-skin AUC by 0.269 and 0.248 using the same images. The information needed to spot dark-skin cancers was in the pictures all along; Deep Derm was not using it.

A model that has already failed on a subgroup cannot be fixed by rescaling at run time, but two different representations were able to recover the signal. These findings do not settle the question for clinical use. Independent, prospective threshold setting and a controlled look at Deep Derm's internal features are still needed.

For fairness audits and dataset selection, the choice of representation matters as much as calibration and threshold setting. When a benchmark classifier seems to have hit an equity ceiling, testing alternative frozen representations on the same data can reveal whether the ceiling comes from the images or from the model[11].

## References

1. Mahendra raj K, Sidhu K, Englesbe MJ. Malignant melanoma in African-Americans. *Medicine*. 2017. 96: 6258.
2. American Cancer Society. Cancer facts & figures 2024. American Cancer Society. <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2024-cancer-facts-figures.html>.
3. Esteva A, Kuprel B, Novoa RA. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017. 542: 115-118.
4. Daneshjou R, Vodrahalli K, Novoa RA. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*. 2022. 8: 6147.
5. Groh M, Harris C, Soenksen L. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2021. 1820-1828.
6. U.S. Food and Drug Administration. Executive summary: General and Plastic Surgery Devices Panel meeting. 2022. <https://www.fda.gov/media/160251/download>
7. Google Health. Derm Foundation model card. Health AI Developer Foundations. <https://health.google/developers/>. 2024.
8. Oquab M, Darcet T, Moutakanni T. DINOv2: Learning robust visual features without supervision. *arXiv*. 2023. <https://arxiv.org/abs/2304.07193>
9. Matsoukas C, Haslum J F, Söderberg M, Smith K. Evaluating general-purpose vision foundation models for

- medical image analysis. arXiv. 2023. <https://arxiv.org/abs/2312.02366>
10. Centers for Disease Control and Prevention. Health and economic benefits of skin cancer interventions. <https://www.cdc.gov/nccdphp/priorities/skin-cancer.html>. 2025.
11. Centers for Disease Control and Prevention. Health and economic benefits of skin cancer interventions. <https://www.cdc.gov/nccdphp/priorities/skin-cancer.html>. 2025.