A Systematic Comparison of Pretrained Visual Representations for Dermatologic Concept Encoding

Linda Wermelinger[1,2], Simone Lionetti[2], Nipun Ranasekara[1,2],
Marc Pouly[2], Alexander A. Navarini[1,3]

[1]University of Basel; [2]Lucerne University of Applied Sciences and Arts;
[3]University Hospital of Basel

Abstract

AI-based decision support systems have limited value for clinicians when they produce predictions without providing interpretable explanations. In dermatology, it is established that deep visual representations can encode clinically meaningful concepts, yet the consistency of this encoding across different model architectures and the degree of entanglement with dataset-specific characteristics remain under-explored. In this work, we present a systematic comparison of visual representations learned by 11 pretrained models, including general-purpose (ViT-B/16), self-supervised (DINOv2/v3), and multimodal architectures (CLIP, SigLIP), as well as those adapted specifically for the dermatological domain (e.g., MedSigLIP, MAKE). Using the SkinCon dataset, we evaluate image embeddings through dimensionality reduction, linear probing for concept-level classification, and distance-based evaluation. While several models achieve strong concept classification performance (MedSigLIP peak AUC-ROC: 0.9266 ± 0.0001), low-dimensional projections often show pronounced grouping by image source rather than dermatological concepts. Our analysis reveals that this clustering can be driven by specific non-clinical artifacts, such as the presence of a measurement ruler, which may overshadow clinical morphology in the representation space. Cross-dataset evaluation on the PH2 dataset indicates that while some visual attributes generalize (e.g., Black AUC-ROC: 0.861), others remain sensitive to differences in image modality and class imbalance. These findings provide a systematic overview of current model capabilities and illustrate the extent to which clinical semantics can remain entangled with dataset-specific features, highlighting an important area for refinement in the development of transparent AI-assisted dermatologic diagnosis.

Introduction

AI-based decision support systems have limited value for clinicians if they solely produce decisions without providing explanations[1]. Explainable artificial intelligence (XAI) addresses this limitation by supporting clinical decision-making with transparent and interpretable reasoning. In dermatology, although vision models are known to encode clinically meaningful concepts, a systematic comparison of how different architectures organize these concepts is lacking[2,19]. Furthermore, the extent to which modern pretrained models capture clinically relevant semantics rather than dataset-specific characteristics is not fully understood[2,19,20,21].

In this work, we analyze the interpretability of deep visual representations for dermatologic diagnosis using the SkinCon dataset, a multilabel clinical image dataset annotated with 31 dermatological concepts (e.g., erythema, plaque, and scale). We evaluate embeddings from a diverse set of pretrained models through dimensionality reduction, concept-level classification, and cross-dataset analysis. Our study provides insights into how different representation learning approaches align with clinically grounded concepts and highlights limitations relevant for explainable AI in dermatology.

Methods

We use the SkinCon dataset[3], a multilabel dermatological image dataset consisting of images manually annotated by two dermatologists with clinically relevant concepts such as lesion morphology and visual attributes. Images are drawn from two clinical sources, DDI[4] and Fitzpatrick17k[5]. To ensure reliable concept-level evaluation, we exclude concepts with fewer than 30 samples. The dataset is further cleaned using CleanPatrick[6] to remove off-topic samples and near duplicates.

Image embeddings are extracted from a diverse set of pretrained deep learning models, including general-purpose vision models, self-supervised foundation models, multimodal image–text models, and dermatology-specific networks. To isolate the impact of training objectives, we primarily compare five models utilizing a consistent ViT-B/16 backbone (ViT[7], DINOv3[8], CLIP[9], PanDerm[10], DermLIP[11], MAKE[12]). For each model, fixed image representations are obtained without additional fine-tuning. We analyze representation structure using UMAP[13] for qualitative inspection and linear probing for concept-level classification, reporting validation AUC-ROC, F1-scores, and Average Precision. To ensure robustness, linear probing was performed over 10 iterations, reporting the mean and standard deviation. . For multimodal models, a distance-based evaluation additionally compares image embeddings to text embeddings of dermatological concepts. To assess generalization, we apply the top-performing self-supervised model (DINOv3) to evaluate representations on the PH2[14] image dataset.

Results

Qualitative inspection of UMAP projections revealed that embeddings from most models did not form distinct clusters according to dermatological concepts. Instead, the embeddings formed clusters aligned with image source, implying that the models encoded dataset-specific traits. Analysis of the DDI source showed that these clusters are primarily driven by

non-clinical artifacts. Specifically, images containing measurement rulers used for sizing lesions formed distinct groups that obscured the clinical features (Figure 1).

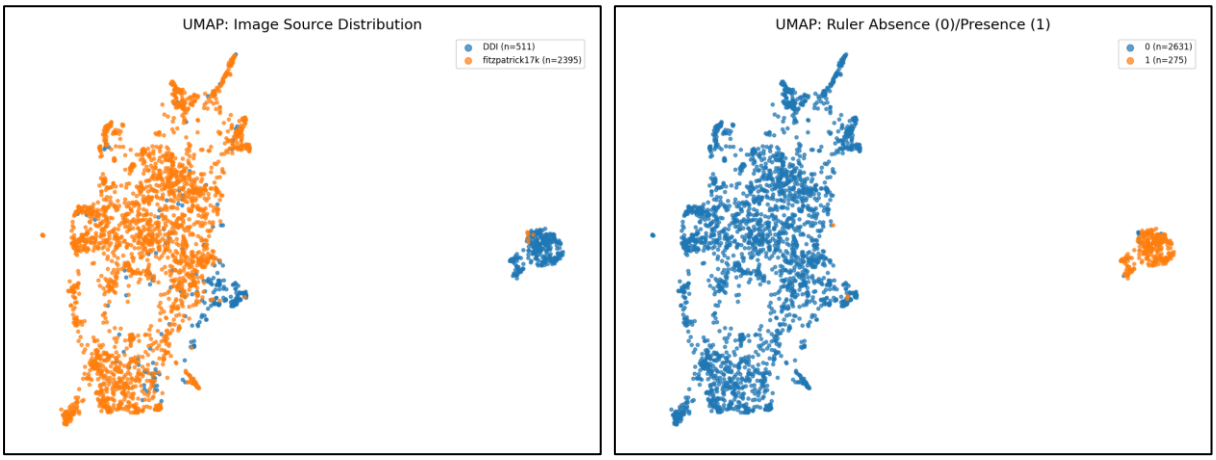

Figure 1. UMAP projections of embeddings from DINOv3. Left: Embeddings colored by image source, showing distinct clustering of different datasets. Right: DDI source images specifically labeled by the presence of measurement rulers. The alignment between the DDI sub-cluster and the ruler label indicates that non-clinical artifacts are a primary driver of the representation structure.

Quantitative concept classification using linear probing showed that performance peaked with MedSigLIP[15] followed by MAKE (Table 1). Our controlled comparison of ViT-B/16 backbones demonstrates that domain-specific pretraining improves the encoding of dermatologic concepts over general-purpose baselines. Common concepts such as Erythema (Max F1: 0.88) were consistently identified, while domain-specific models enabled the detection of subtle features like Pustule (F1: 0.69 MedSigLIP vs. 0.00 ViT-B/16 baseline). Distance-based evaluation of multimodal models exhibited a similar pattern, with some concepts more accurately retrieved through image–text similarity than others. Comparisons between DINOv2 and DINOv3 revealed an improvement (+0.027 AUC-ROC) in the newer version.

Table 1. Comparison of concept classification with linear probing on SkinCon.

	Vision Encoder	Training Paradigm	Val AUC-ROC (Mean ± SD)
ViT-B/16 (baseline)	ViT-B/16	Supervised (ImageNet)	0.7771±0.0013
DINOv3	ViT-B/16, LVD-1689M	Self-Supervised	0.8912±0.0004
DINOv2[16]	ViT-B/14	Self-Supervised	0.8639±0.0014
Derm Foundation[17]	ResNet101x3	Dermatology-Specific	0.8837±0.0005
PanDerm	ViT-B/16	Dermatology-Specific	0.8850±0.0001
CLIP	ViT-B/16	Multimodal	0.8769±0.0003
MONET[2]	ViT-L/14	Dermatology-Specific	0.8822±0.0003
MAKE	ViT-B/16	Dermatology-Specific	**0.9173±0.0003**
DermLIP	ViT-B/16	Dermatology-Specific	0.9140±0.0003
SigLIP[18]	ViT-L/16	Multimodal	0.9147±0.0003
MedSigLIP	So400m	Dermatology-Specific	**0.9266±0.0001**

Using DINOv3, we evaluated the transferability of representations to the PH2 dataset. Results show that some visual attributes generalize across datasets (Black AUC-ROC: 0.861), while others are sensitive to differences in image modality and class imbalance (e.g., Erythema AUC-ROC: 0.385 at 5% prevalence), reflecting the challenges of domain shift between clinical and dermatoscopic images.

Discussion and Conclusions

Our systematic comparison demonstrates that while modern vision models successfully encode dermatologic concepts, these representations remain sensitive to domain shifts and low-level artifacts like rulers. This underscores the importance of evaluating quantitative performance and qualitative structure to ensure that XAI systems rely on clinically grounded features rather than biases.

References

1. Chanda T, Hauser K, Hobelsberger S, Bucher TC, Garcia CN, Wies C, et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications*. 2024 Jan 15;15(1):524. doi:10.1038/s41467-023-43095-4.
2. Kim C, Gadgil SU, DeGrave AJ, Omiye, JA, Cai ZR, Daneshjou R, et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature Medicine*. 2024 April 16;30:1154–1165. doi:10.1038/s41591-024-02887-x
3. Daneshjou R, Yuksekgonul M, Cai ZR, Novoa R, Zou JY. SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Advances in Neural Information Processing Systems*. 2022 Dec 6;35:18157–67.
4. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*. 2022 Aug 12;8(32).
5. Groh M, Harris C, Soenksen L, Scale F, Francisco S, Scale R, et al. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*; 2021. p. 1820–1828.
6. Gröger F, Lionetti S, Gottfrois P, Gonzalez-Jimenez A, Amruthalingam L, Goessinger EV, et al. CleanPatrick: a benchmark for image data cleaning. *arXiv.org*. 2025. doi:10.48550/arXiv.2505.11034.
7. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. *ICLR*. 2021.
8. Siméoni O, Vo HV, Seitzer M, Baldassarre F, Oquab M, Jose C, et al. Dinov3: self-supervised learning for vision at unprecedented scale. *arXiv.org*. 2025. doi:10.48550/arXiv.25.08.10104.
9. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. *arXiv.org*. 2021. doi:10.48550/arXiv.2103.00020
10. Yan S, Yu Z, Primiero C, Vico-Alonso C, Wang Z, Yang L, et al. A multimodal vision foundation model for clinical dermatology. *Nat Med*. 2025 Jun 6;1–12.
11. Yan S, Hu M, Jiang Y, Li X, Fei H, Tschandl P, et al. Derm1M: a million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. *arXiv.org*. 2025. doi:10.48550/arXiv.2503.14911.
12. Yan S, Li X, Hu M, Jiang Y, Yu Z, Ge Z. MAKE: multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment. MICCAI. 2025.
13. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv.org*. 2018. doi:10.48550/arXiv.1802.03426
14. Mendonca T, Ferreira PM, Marques JS, Marcal ARS, Rozeira J. PH² - a dermoscopic image database for research and benchmarking. *Annu Int Conf IEEE Eng Med Biol Soc*. 2013;2013:5437–40.
15. Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, et al. MedGemma technical report. *arXiv.org*. 2025. doi:10.48550/arXiv.2507.05201.
16. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: learning robust visual features without supervision. *arXiv.org*. 2024. doi:10.48550/arXiv.2304.07193.
17. Google for Developers. Derm foundation model. 2025. Available from: https://developers.google.com/health-ai-developer-foundations/derm-foundation.
18. Zhai X, Mustafa B, Kolesnikov A, Beyer L. Sigmoid loss for language image pre-training. ICCV. 2023. p. 11941–52. doi:10.1109/ICCV51070.2023.01100.
19. Hauser K, Kurz A, Haggenmüller S, Maron RC, von Kalle C, Utikal JS, et al. Explainable artificial intelligence in skin cancer recognition: a systematic review. *European Journal of Cancer*. 2022 May 1;167:54–69. doi:10.1016/j.ejca.2022.02.025.
20. Wang WC, Ahn E, Feng D, Kim J. A review of predictive and contrastive self-supervised learning for medical images. *Mach Intell Res*. 2023 Aug 1;20(4):483–513. doi:10.1007/s11633-022-1406-4.
21. Lu Y, Zając HD, Cheplygina V, Jiménez-Sánchez A. Intuitions of machine learning researchers about transfer learning for medical image classification. *arXiv.org*. 2025. doi:10.48550/arXiv.2510.00902.

This research was funded by the Swiss National Science Foundation (SNSF) under grant 20HW-1 228541.