A Framework for Evaluating the Efficacy of Foundation Embedding Models in Healthcare

Sonnet Xu,[1] Haiwen Gui*,[1] Veronica Rotemberg,[2] Tongzhou Wang,[3] Yiqun T. Chen**,[1] Roxana Daneshjou**[1]*

[1]Stanford University, School of Medicine, Stanford, CA [2]Memorial Sloan Kettering Cancer Center, New York, NY [3]Massachusets Institute of Technology, Cambridge, MA, USA

**Both authors contributed equally to this work **co-senior authorship*

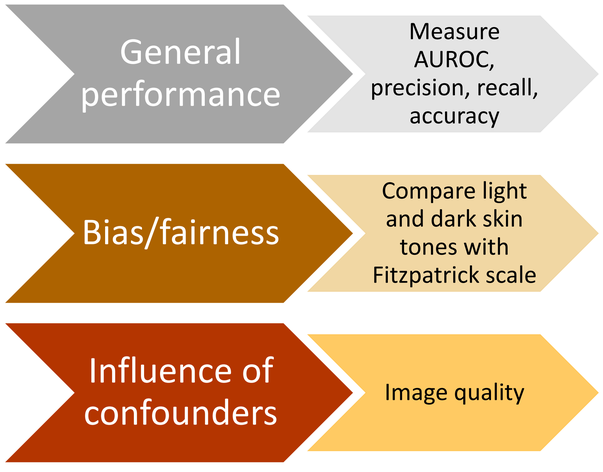
Stanford Medicine

Background

- Accessible healthcare can be enabled by artificial intelligence tools
- Current methods such as the Convolutional Neural Network (CNN) are computationally expensive to train and rely on large amounts of domain-specific data
- Foundation models offer the potential to reduce the data, computing, and technical expertise necessary for building models from scratch
- Bias and other concerns limits their applicability

Purpose

- Develop a framework for assessing the foundation model
- Used Google's newly released Derm Foundation Embedding model as an example in dermatology



Methods

- Generated 656 Diverse Dermatology Image (DDI) embeddings and 10,015 ISIC embeddings
- Built a logistic regression classifier on top of the embeddings to assess accuracy
- Compared Derm Foundation embeddings with MONET embeddings
- Computed cosine similarities between benign and malignant skin conditions across different skin tones
- used ℓ2-regularized logistic regressions to assess the test set performance by training on high or low-quality images only

Results

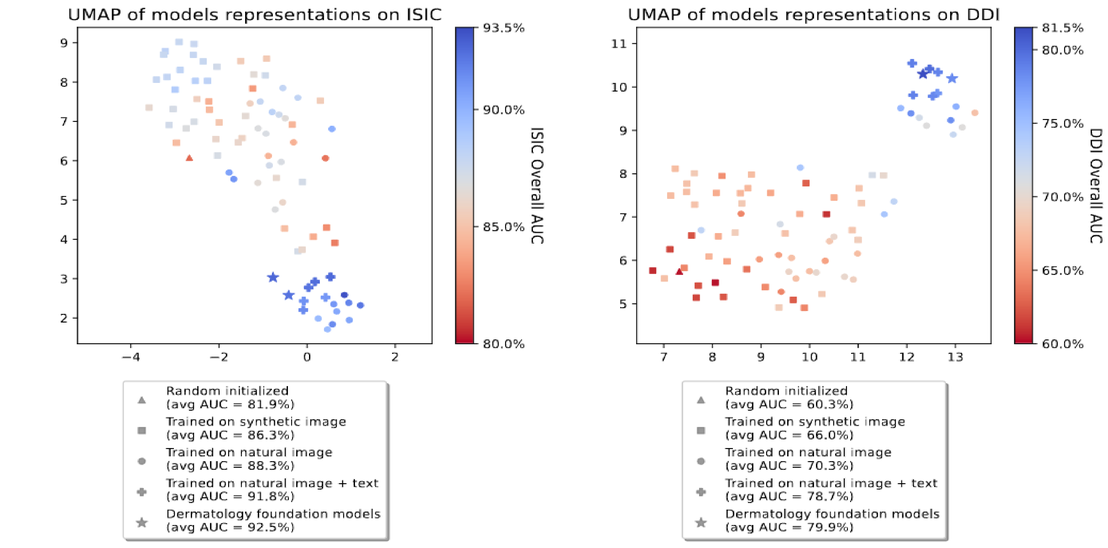
Classifier Accuracy

- Derm Foundation embeddings with a simple logistic classifier produced competitive results to SOTA.
- Models trained with only natural images or synthetic images perform much worse on the downstream prediction tasks

The gap between dermatology foundation models and general-purpose CLIP foundation models is quite small (< 2% AUC).

Dataset	Class	Specificity	Precision	Recall	F1-score	AUROC	Accuracy
DDI	Malignancy	0.94	0.82	0.58	0.68	0.76	0.83
ISIC 2018	Malignancy	0.94	0.71	0.61	0.66	0.78	0.88
ISIC 2018	akiec	0.99	0.65	0.53	0.58	0.76	0.53
ISIC 2018	bcc	0.99	0.75	0.67	0.70	0.83	0.67
ISIC 2018	bkl	0.96	0.70	0.67	0.69	0.82	0.67
ISIC 2018	df	1.00	0.94	0.65	0.77	0.83	0.65
ISIC 2018	mel	0.96	0.66	0.54	0.59	0.75	0.54
ISIC 2018	nv	0.80	0.90	0.96	0.93	0.88	0.96
ISIC 2018	vasc	1.00	0.84	0.91	0.87	0.96	0.91

Table 1: Predictive performance of using an ℓ_2 logistic regression on Derm Foundation embeddings to predict malignant versus benign lesions on DDI (first row) and ISIC (second row), as well as performance on predicting specific diagnoses (one-versus-rest) on ISIC (third to last rows).



Skin Tone and Data Quality

The Fitzpatrick Skin Tone (FST) dominates the cosine similarity calculation.

Given samples from the same FST class, benign samples are more similar than malignant samples.

Training on high-quality images improves the test performance on low-quality images by a large margin (around 0.2 improvement across precision, recall, F1, and AUROC) compared to training and testing on low-quality images.

Train	Test	Specificity	Precision	Recall	F1	AUC	Accuracy
I/II[a]	I/II	0.88 ± 0.05	0.56 ± 0.12	0.49 ± 0.17	0.69 ± 0.08	0.78 ± 0.07	0.79 ± 0.06
III/IV	I/II	0.72	0.41	0.63	0.64	0.76	0.70
V/VI	I/II	0.93	0.54	0.27	0.61	0.68	0.77
III/IV[a]	III/IV	0.89 ± 0.04	0.73 ± 0.06	0.62 ± 0.11	0.76 ± 0.03	0.83 ± 0.05	0.81 ± 0.02
I/II	III/IV	0.84	0.65	0.66	0.75	0.82	0.79
V/VI	III/IV	0.94	0.67	0.27	0.61	0.69	0.73
V/VI[a]	V/VI	0.93 ± 0.02	0.72 ± 0.06	0.56 ± 0.09	0.76 ± 0.04	0.82 ± 0.04	0.85 ± 0.02
I/II	V/VI	0.73	0.36	0.50	0.60	0.68	0.68
III/IV	V/VI	0.82	0.37	0.35	0.59	0.68	0.71

Table 2: Predictive performance of using an ℓ_2 logistic regression on Derm Foundation

FST Categories	Classes	Average Cos Sim
I/II & I/II	benign & benign	0.55
I/II & I/II	benign & malignant	0.50
I/II & I/II	malignant & malignant	0.53
I/II & V/VI	benign & benign	0.37
I/II & V/VI	malignant & malignant	0.32
V/VI & V/VI	benign & benign	0.48
V/VI & V/VI	benign & malignant	0.41
V/VI & V/VI	malignant & malignant	0.43

Table 3: Average cosine similarities between Derm Foundation embeddings across different diagnostic classes and Fitzpatrick Skin Tones.

Conclusions and Discussion

- Derm Foundation and MONET are more accurate than previous SOTA, indicating foundation models have potential for medical applications.
- General-purpose CLIP models should be more broadly considered as a baseline for medicine and domain-specific models should be compared to and developed based on these performant general-purpose models.
- Higher cosine similarities between same FST category, than between same classification label suggests that machine learning models implicitly factor skin tone into internal evaluations.
- Models trained on high-quality images generally outperform those trained on low-quality images.

Limitations and future work

- Ambiguous input data for closed-source models raises concerns about potential data leakage
- Constraints presented by limited availability of labeled skin color datasets
- Expand beyond dermoscopy images and teledermatology photos to other modalities
- The generalizability of our findings needs to be validated in other medical domains and applications
- Develop a comprehensive suite of datasets and models for evaluating medical AI models
- Incorporate multimodal medical data beyond images, such as patient metadata and history
- Extend and assess this framework in other medical fields such as cardiology, radiology, and pathology, where there is an abundance of multimodal data and foundation models are beginning to show promise

Key References

Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). CoRR, abs/1902.03368, 2019. URL http://arxiv.org/abs/1902.03368.

R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. JAMA Dermatol, 157(11):1362–1369, Nov 2021.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. Disparities in dermatology ai performance on a diverse, curated clinical image set. Science Advances, 8(32):eabq6147, 2022. doi: 10.1126/sciadv.abq6147. URL https://www.science.org/doi/abs/10.1126/sciadv.abq6147.

T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. Arch Dermatol, 124(6):869–871, Jun 1988

H. Gui, J. A. Omiye, C. T. Chang, and R. Daneshjou. The Promises and Perils of Foundation Models in Dermatology. J Invest Dermatol, Mar 2024