Bridging Research and Practice: A Systematic Evaluation of Generalist and Dermatology-Specific Models in Clinical Skin Lesion Classification

Emanoel dos Santos[1], Kelvin Cunha[1], Rodrigo Mota[1], Fabio Papais[1], Thales Bezerra[1], Natalia Lopes[1], Erico Medeiros[1], Shirley Cruz[2], Jessica Araujo[2], Paulo Borba[1], and Tsang Ing Ren[1]

[1] Centro de Informática, Universidade Federal de Pernambuco (UFPE), Brazil
{etcs, kbc, raafm, flfp, tob2, nal, emm2, phmb, tir}@cin.ufpe.br
[2] Hospital das Clinicas, Ebserh, UFPE, Pernambuco, Brazil

Abstract. The application of machine learning to dermatology has grown substantially in recent years, moving beyond proof-of-concept studies toward potential applications. However, clinical dermatology remains a challenging and still open problem. Diagnostic assessment is often ambiguous, and skin lesions exhibit high variability, compounded by differences in acquisition modality, device quality, and patient demographics. These factors hinder the development of robust models suitable for safe and equitable clinical use. To support translation into practice, it is essential to systematically evaluate how contemporary models generalize across heterogeneous data sources. In this work, we benchmark a diverse set of architectures on recent dermatology datasets, spanning dermoscopic images and smartphone-based clinical photographs. We assess the robustness of recent general-purpose and medical vision-language models, as well as foundation models, and compare them against task-specific dermatology classifiers, including embedding-based approaches and convolutional neural networks. Our study provides an evaluation of model performance under distribution shifts, modality changes, and demographic variability. By quantifying the gap between current state-of-the-art models and the requirements of clinical deployment, we aim to contribute to the development of reliable, accessible, and clinically applicable AI systems for dermatology.

Keywords: Clinical Dermatology · Vision language models · Task-specific Classification.

1 Introduction

Skin cancer is among the most common malignancies worldwide, and timely recognition is essential on reducing patient risk of developing serious conditions [2]. Clinical practice diagnosis relies on visual assessment through dermoscopy or standard photography, often under heterogeneous acquisition con-

ditions [9]. Also, operator expertise and patient characteristics introduce substantial variability, making consistent and reliable assessment challenging [4,12]. Any artificial intelligence system intended for clinical use must therefore demonstrate robustness across modalities and real-world data conditions. Recent advances in large-scale foundation models and vision-language models (VLMs) have generated considerable enthusiasm in medical imaging [14,18]. Their broad pretraining and multimodal capabilities suggest improved generalization and contextual reasoning compared to traditional task-specific convolutional networks. As a result, increasingly large models are being adopted for dermatology tasks, sometimes with limited domain-specific adaptation. However, high performance on curated benchmarks does not guarantee reliability across imaging modalities, datasets, or patient populations. Clinical translation requires systematic validation under realistic variability. Models must be evaluated across datasets collected under different protocols, and under distributions that reflect actual deployment scenarios. Without such evaluation, it is unclear whether newer large-scale architectures meaningfully outperform specialized dermatology models or whether they introduce new failures. In this work, we conduct a structured benchmark of diverse modeling paradigms for binary malignancy risk prediction across multiple dermatology datasets. We compare dermatology-specific embedding models, conventional convolutional networks, general representation learners, vision-language models, medical-adapted VLMs, and multimodal approaches incorporating metadata. Our evaluation spans modality-specific training, merged datasets, and cross-dataset testing. This study provides empirical evidence on the extent to which recent architectural advances translate into clinically meaningful robustness. Code is available at `https://github.com/TIC-13/derm-bench`.

2 Methodology

2.1 Problem Formulation

We address binary malignancy risk prediction from dermatological images. Given an input image x, the task is to predict a label $y \in \{0, 1\}$, where $y = 1$ denotes malignant and $y = 0$ benign lesions. The objective is to evaluate model robustness across datasets and merged training configurations. Models were evaluated across multiple publicly available dermatology datasets encompassing dermoscopic and smartphone-based images. To assess robustness under heterogeneous conditions, we defined the following evaluation: (i) individual datasets, (ii) merging datasets by modality (clinical or dermoscopic), (iii) and merging all datasets. For merged settings, datasets were combined at the image level while preserving original labels. To ensure fair comparison across model families identical train/test splits were used across models adopting the 70/15/15 split configuration, except where splits are already provided by the original source. Image preprocessing (resizing and normalization) was standardized and no test data leakage across datasets was allowed.

2.2 Model Families

We evaluated three model families reflecting current approaches in dermatological AI:

- **Dermatology-Specific and General Embedding Models**: We evaluated domain-specific and general embedding backbones, including the derm-foundation [10,21] and the self-supervised DinoV3 [15]. For these models, the backbone was used as a frozen feature extractor and image embeddings were computed for all samples. A downstream classifier was trained on top of the embeddings (i.e, MLP, SVM, Random Forest, and XGBoost). This design decouples representation learning from classification and allows consistent comparison across different backbone models.
- **Convolutional and Vision Transformer Models**: We evaluated task-specific supervised architectures trained end-to-end, including ConvNeXt, EfficientNet, ResNet, and Vision Transformer (ViT). Models were fine-tuned on each dataset using cross-entropy loss.
- **Vision Language Models (VLMs)**: Large multimodal models were evaluated using prompt-based inference strategies, including: Dual-encoder configurations, simple prompts, structured prompts based on dermatological criteria, prompts augmented with patient metadata when available. We tested the models: BiomedCLIP [20], CLIP ViT-B/32, CLIP ViT-L/14 [13], MedSigLIP-448 [19], Gemma-3 27B, LLaMA-3.2-Vision 11B, LLaMA-3.2-Vision 90B, LLaMA-4 (16×17B) [1], Qwen-2.5VL 72B [3].

2.3 VLMs variations

We evaluated multiple prompt and inference configurations to assess the behavior of large VLMs in dermatological malignancy prediction.

- **Simple Prompt (Zero-Shot Classification)**: In the baseline configuration, the VLM receives the image and a direct instruction to classify the lesion as *benign* or *malignant*. The system prompt defines the model as an expert dermatologist, while the user prompt restricts the output to a single-word label. This setup evaluates pure zero-shot visual-language reasoning without explicit clinical heuristics.
- **ABCDE-Guided Prompt**: To incorporate structured dermatological knowledge, we augmented the system prompt with an internal checklist based on the ABCDE criteria [7] and the "ugly duckling" sign [8]. The model is instructed to apply these warning signs silently when making its decision, while still outputting only a single label. This configuration tests whether embedding domain heuristics into the prompt improves classification consistency and alignment with clinical reasoning.
- **Metadata-Augmented Prompt**: In this variant, structured clinical metadata (when available) is provided alongside the image. The system prompt explicitly instructs the model to integrate both visual and clinical information in its decision. This configuration evaluates whether contextual patient information enhances malignancy prediction.

– **Dual-Encoder Classification**: In the dual-encoder setup, the model processes image and label representations separately and predicts the most likely class via similarity matching. Rather than generating free-text responses, the model selects between predefined class labels (*benign*, *malignant*). This approach evaluates discriminative multimodal alignment without reliance on generative prompting.

2.4 Datasets

We evaluated our models on publicly available dermatology datasets spanning dermoscopic and clinical imaging modalities. Dermoscopic benchmarks include HAM10000 [17],ISIC 2018 [5], and ISIC 2024 [11]. Clinical datasets include PAD-UFES-20 (PAD) [12], which provides smartphone images with structured patient metadata, HC [4], reflecting heterogeneous real-world acquisition conditions, DDI [6], designed to improve skin tone diversity and fairness evaluation, and SD-198 [16], covering a broad range of skin diseases. To assess robustness under heterogeneous training conditions, we additionally constructed merged datasets combining sources: MERGED-ALL (all modalities), MERGED-CLINIC (clinical only), and MERGED-DERM (dermoscopic only), simulating multi-institutional deployment scenarios.

2.5 Class Imbalance and metrics

Given the inherent class imbalance in dermatology datasets, we evaluated three training strategies (i) Unbalanced training, (ii) Random undersampling, (iii) SMOTE-based oversampling. For embedding based pipelines, imbalance handling was applied during classifier training. For end-to-end CNN/ViT models, imbalance handling was implemented at the data sampling stage. The best-performing configuration per dataset was used for family-level comparison. Performance was assessed using standard metrics, as in the source datasets, we focus on the F1-Score to discuss our analyzes due to class imbalance and the clinical importance of balancing false positives and false negatives in malignancy risk prediction.

3 Results

3.1 Overall Performance Across Datasets

We first compared the best-performing configuration of each model family on each dataset. Table 1 summarizes the highest F1 achieved per family and dataset. Embedding-based models achieved the highest average score (88.6%), followed by CNN/ViT models (86.6%), while VLMs achieved substantially lower performance (65.8%). Thus, VLMs consistently ranked last across all evaluation settings. The performance gap between VLMs and the best vision-only approaches

Table 1. Best Macro F1-score per model family across datasets.

Dataset	VLM	CNN/ViT	Embeddings
PAD	0.7671	0.9162	0.9343
HC	0.6123	0.8281	0.8760
ISIC18	0.6026	0.8509	0.8738
ISIC24	0.7048	0.8444	0.8333
DDI	0.6667	0.9112	0.8712
HAM10000	0.5622	0.8859	0.9043
SD-198	0.8426	0.8552	0.9638
MERGED-ALL	0.6114	0.8664	0.8783
MERGED-CLINIC	0.6375	0.8624	0.8888
MERGED-DERM	0.5794	0.8384	0.8503
Average F1	**0.6587**	**0.8659**	**0.8874**

are substantial. In malignancy risk prediction, it may translate into a considerable increase in false negatives or false positives, directly affecting triage decisions. The persistence of this gap across dermoscopic, clinical, and merged settings suggests that model scale and multimodal pretraining alone do not ensure clinically reliable performance in dermatology.

When grouping datasets by modality, embedding and CNN/ViT models demonstrated stable performance across both dermoscopic and clinical images, supporting their robustness under varying acquisition conditions. In contrast, VLM performance showed higher variability and did not demonstrate consistent advantages on clinical photographs despite incorporating language-based reasoning. Notably, dermatology-specific embeddings achieved particularly strong performance on SD-198, indicating that domain-aligned representations can effectively capture clinically lesion patterns.

In the merged training configurations (MERGED-ALL, MERGED-CLINIC, MERGED-DERM), the relative ranking of model families remained unchanged. Embedding-based approaches maintained the highest average performance ($F1 \approx 88\%$), followed by CNN/ViT models ($F1 \approx 86\%$), while VLMs remained substantially lower ($F1 \approx 65\%$). This suggests that increasing dataset diversity alone does not compensate for architectural misalignment with the task, and that domain-specific visual representations remain critical for consistent malignancy assessment.

3.2 Impact of Class Imbalance

To assess the influence of imbalance mitigation strategies, we compared the best-performing embedding-based configuration under three training settings (unbalanced, random undersampling, and SMOTE). Table 2 reports the results obtained under each strategy. Overall, unbalanced training achieved the highest scores, followed by SMOTE and undersampling. While imbalance handling did not drastically alter performance in most datasets, certain settings exhibited notable sensitivity. For instance, undersampling improved performance on DDI

Table 2. Best Macro F1-score achieved under different class imbalance strategies (embedding-based models).

Dataset	Unbalanced	Undersampling	SMOTE
PAD	0.9343	0.9124	0.9152
HC	0.8760	0.8680	0.8567
ISIC18	0.8738	0.7797	0.8716
ISIC24	0.8313	0.8333	0.7955
DDI	0.8224	0.8712	0.8071
HAM10000	0.9022	0.8050	0.9043
SD-198	0.9638	0.9638	0.9396
MERGED-ALL	0.8771	0.8583	0.8783
MERGED-CLINIC	0.8844	0.8872	0.8888
MERGED-DERM	0.8503	0.8069	0.8412
Average F1	**0.8816**	**0.8586**	**0.8698**

Table 3. Average Macro Best F1-score per architecture across evaluation categories.

Architecture	Dermoscopic	Clinical	Merged-All
Derm-Foundation	0.8308	0.8872	0.8714
DINOv3 ViT-7B/16	0.8503	0.8888	0.8783
DINOv2 Giant	0.8383	0.8479	0.8641
ConvNeXt (best)	0.8373	0.8624	0.8619
EfficientNet (best)	0.8171	0.8218	-
Best VLM (prompt-based)	0.5794	0.6723	0.6114

(87% vs. 82% unbalanced), whereas SMOTE led to performance degradation in ISIC24 and SD-198. These findings highlight that class rebalancing strategies must be carefully validated, as synthetic oversampling or aggressive undersampling may inadvertently affect sensitivity to malignant cases. Robust evaluation across imbalance settings is therefore essential before deployment in screening or triage scenarios.

3.3 Architectures validation

To provide a more granular view of performance, we summarize results at the architecture level rather than by model family. We report the average of the best F1-scores obtained per dataset, grouped into three clinically relevant categories: dermoscopic datasets (ISIC18, ISIC24, HAM10000, MERGED-DERM), clinical datasets (PAD, HC, DDI, SD-198, MERGED-CLINIC), and fully merged datasets (MERGED-ALL).

Across dermoscopic datasets, large vision backbones (EfficientNet, DINOv3, Derm-Foundation) achieved comparable performance. In clinical datasets, Derm-Foundation and EfficientNet showed the strongest robustness, while VLMs remained substantially lower. In the merged-all setting, embedding-based architectures maintained stable performance, whereas VLMs did not exhibit improvement despite increased dataset diversity. These results indicate that per-

Table 4. F1-score comparison of VLM variants on datasets with available patient metadata.

Setting	Model	HC	PAD	ISIC18	ISIC24
Dual Encoder	BiomedCLIP	0.3362	0.3951	0.3375	0.3041
	CLIP-B/32	0.5882	0.4866	0.5185	0.5382
	CLIP-L/14	0.3191	0.4866	0.1740	0.6116
	MedSigLIP	0.5686	0.6387	0.4815	0.4755
Prompt + ABCDE	Gemma-3 27B	0.5313	0.5504	0.2040	0.7048
	LLaMA-3.2V 11B	0.3087	0.4090	0.1367	0.2823
	LLaMA-3.2V 90B	0.4962	0.5217	0.3725	0.5980
	LLaMA-4 16×17B	0.6123	0.6476	0.1632	0.4494
	Qwen-2.5VL 72B	0.5175	0.6213	0.4243	0.6088
Prompt + Metadata	Gemma-3 27B	0.4693	0.7381	0.3203	0.6127
	LLaMA-3.2V 11B	0.4756	0.4290	0.2010	0.2823
	LLaMA-3.2V 90B	0.3003	0.4398	0.2174	0.3650
	LLaMA-4 16×17B	0.3099	0.4141	0.2482	0.4168
	Qwen-2.5VL 72B	0.4693	0.7671	0.6026	0.5507
Prompt (Simple)	Gemma-3 27B	0.4930	0.5685	0.3716	0.6451
	LLaMA-3.2V 11B	0.5357	0.6835	0.5524	0.4744
	LLaMA-3.2V 90B	0.4356	0.3662	0.4197	0.4457
	LLaMA-4 16×17B	0.3454	0.3483	0.3093	0.3967
	Qwen-2.5VL 72B	0.4872	0.5035	0.5680	0.5644

formance differences are primarily driven by representation quality rather than model scale alone. Architectures trained or aligned with dermatological visual features demonstrate greater stability across acquisition modalities. Notably, DINO-based architectures achieved performance comparable to dermatology-specific trained models, despite relying on general self-supervised feature representations. Fine-tuning strategies further improved their results, highlighting the strength of transferable visual representations when properly adapted to the task.

3.4 VLMs Variations

For this test, we report results only on datasets for which structured patient metadata is available (HC, PAD, ISIC18, ISIC24), enabling a fair comparison across prompt-based and metadata-augmented VLM configurations.

Dual-encoder models exhibit more stable behavior than prompt-based VLMs, suggesting that discriminative image-text alignment is less sensitive to prompt design than generative reasoning. Among them, MedSigLIP and CLIP-B/32 consistently outperform BiomedCLIP, despite the latter being trained on biomedical data. Prompt engineering using ABCDE heuristics leads to moderate gains in selected cases (e.g., ISIC24), but improvements are inconsistent and highly model-dependent. Incorporating patient metadata yields the strongest improvements for a subset of models, particularly on PAD, indicating that contextual clinical information can partially compensate for visual ambiguity.

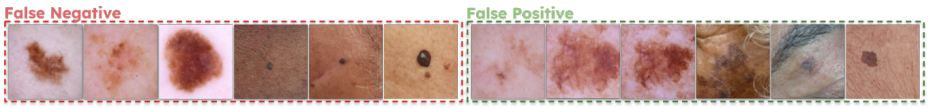

Fig. 1. Qualitative false negative (left) and false positive (right) cases in malignancy prediction on dermoscopic (HAM10000) and clinical (HC) data.

Overall, these results suggest that current VLMs lack the robustness required for clinical deployment. Performance appears constrained less by prompt sophistication or model scale and more by the intrinsic difficulty of prediction from images alone. Simply applying general-purpose VLMs in zero-shot settings, or augmenting them with additional contextual information, is insufficient to achieve performance comparable to specialized models. Rather than reflecting an inherent limitation of VLMs, these findings highlight the importance of task-specific adaptation and proper training. Although general VLMs may appear to solve the target task, deploying them without proper domain-specific adaptation can pose substantial risks to application reliability.

3.5 Qualitative Analysis

Figure 1 illustrates representative false positive and negative cases. Several benign lesions exhibit irregular borders, color heterogeneity, and partial asymmetry, while some malignant cases present relatively symmetric shapes and limited color variation. This visual overlap highlights a fundamental limitation of image-only malignancy prediction: morphological features alone may be insufficient to reliably separate benign from malignant lesions in all cases. In clinical practice, dermatological assessment extends beyond visual inspection. Patient age, lesion location, temporal evolution, symptoms, and prior history contribute substantially to risk estimation. The absence of such contextual information may explain persistent errors even in high-capacity models. Purely visual classification, particularly under binary malignancy labels, imposes intrinsically ambiguous decision boundaries that cannot be fully resolved through model scale alone.

4 Conclusion

We presented an evaluation of contemporary architectures for binary malignancy risk prediction across heterogeneous dermatology datasets. Our analysis shows that dermatology-aligned embedding models and strong vision backbones consistently outperform large vision-language models. Increasing architectural scale or incorporating multimodal pretraining did not translate into improved clinical robustness. Instead, architectures with domain-aligned visual representations demonstrated the most stable performance under dataset variability. These findings emphasize that rigorous cross-dataset validation and task-specific representation learning remain essential prerequisites for reliable clinical deployment of AI systems in dermatology.

Acknowledgments. The project was supported by the Ministry of Science, Technology, and Innovation of Brazil, with resources from Law No. 8,248, dated October 23, 1991, under the scope of the PPI-SOFTEX, coordinated by Softex and published under RESIDÊNCIA EM TIC 63 – ROBÓTICA E IA – FASE II, DOU 23076.043130/2025-27. For partially supporting this work, we would like to thank INES.IA (National Institute of Science and Technology for Software Engineering Based on and for Artificial Intelligence) `www.ines.org.br`, CNPq grant 408817/2024-0.

Disclosure of Interests

The authors declare that they have no competing interests.

References

1. Abdullah, A.A., Zubiaga, A., Mirjalili, S., Gandomi, A.H., Daneshfar, F., Amini, M., Mohammed, A.S., Veisi, H.: Evolution of meta's llama models and parameter-efficient fine-tuning of large language models: a survey. arXiv preprint arXiv:2510.12178 (2025)
2. Arnold, M., Singh, D., Laversanne, M., Vignat, J., Vaccarella, S., Meheus, F., Cust, A.E., De Vries, E., Whiteman, D.C., Bray, F.: Global burden of cutaneous melanoma in 2020 and projections to 2040. JAMA dermatology **158**(5), 495–503 (2022)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Bezerra, T., Thyago, E., Cunha, K., Abreu, R., Papais, F., Mauro, F., Lopes, N., Medeiros, É., Guido, J., Cruz, S., et al.: Dermai: Clinical dermatology acquisition through quality-driven image collection for ai classification in mobile. arXiv preprint arXiv:2511.10367 (2025)
5. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
6. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J.A.C., Okata-Karigane, U., Zou, J., Chiou, A.S.: Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci. Adv. **8**(32), eabq6147 (Aug 2022)
7. Duarte, A.F., Sousa-Pinto, B., Azevedo, L.F., Barros, A.M., Puig, S., Malvehy, J., Haneke, E., Correia, O.: Clinical abcde rule for early melanoma detection. European Journal of Dermatology **31**(6), 771–778 (2021)
8. Gaudy-Marqueste, C., Wazaefi, Y., Bruneu, Y., Triller, R., Thomas, L., Pellacani, G., Malvehy, J., Avril, M.F., Monestier, S., Richard, M.A., et al.: Ugly duckling sign as a major factor of efficiency in melanoma detection. JAMA dermatology **153**(4), 279–284 (2017)
9. Hasan, M.K., Ahamad, M.A., Yap, C.H., Yang, G.: A survey, review, and future trends of skin lesion segmentation and classification. Computers in Biology and Medicine p. 106624 (2023)

10. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: European conference on computer vision. pp. 491–507. Springer (2020)
11. Kurtansky, N.R., D'Alessandro, B.M., Gillis, M.C., Betz-Stablein, B., Cerminara, S.E., Garcia, R., Girundi, M.A., Goessinger, E.V., Gottfrois, P., Guitera, P., et al.: The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection. Scientific Data **11**(1), 884 (2024)
12. Pacheco, A.G.C., Lima, G.R., Salomão, A.S., Krohling, B., Biral, I.P., de Angelo, G.G., Alves, Jr, F.C.R., Esgario, J.G.M., Simora, A.C., Castro, P.B.C., Rodrigues, F.B., Frasson, P.H.L., Krohling, R.A., Knidel, H., Santos, M.C.S., do Espírito Santo, R.B., Macedo, T.L.S.G., Canuto, T.R.P., de Barros, L.F.S.: PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data Brief **32**(106221), 106221 (Oct 2020)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
14. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
15. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
16. Sun, X., Yang, J., Sun, M., Wang, K.: A benchmark for automatic visual classification of clinical skin disease images. In: European conference on computer vision. pp. 206–222. Springer (2016)
17. 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), 1–9 (2018)
18. Yuceyalcin, F., Yilmaz, A., Temelkuran, B.: A hierarchical benchmark of foundation models for dermatology. arXiv preprint arXiv:2601.12382 (2026)
19. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
20. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
21. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine learning for healthcare conference. pp. 2–25. PMLR (2022)