Taxonomy-Guided Vision–Language Representation Learning for Hierarchical Skin Disease Classification

Anjali Deepu[1] and Swarnava Dey[1⋆]

Indian Institute of Technology Kharagpur, Kharagpur, India
swarnava.dey@kgpian.iitkgp.ac.in

Abstract. Hierarchical skin disease classification must recognize broad disease categories while preserving distinctions among clinically similar subclasses. Parent-level contrastive supervision can work against this goal: sibling diagnoses sharing the same parent are treated as positives, creating a coarse–fine conflict that limits fine-grained discrimination. We introduce a taxonomy-guided vision–language framework that uses language not merely as an input or output modality, but as structured supervision for learning a visual classifier. Multiple concept-derived captions constructed from lesion morphology, anatomical location, symptoms, and disease descriptors provide clinically grounded multi-positive supervision, while same-parent/different-subclass pairs receive increased negative importance according to the disease taxonomy. The resulting representation is coupled with a gallery-based prototype classifier for hierarchy-constrained prediction and retrieval of clinically grounded caption exemplars. Because classes are represented by textual prototypes rather than a fixed output head, the frozen encoders can also operate over a disjoint disease vocabulary through gallery replacement. Taxonomy-guided optimization raises Sub-Class macro-F1 from 14.05 to 32.01, while the complete framework reaches 54.21 and substantially outperforms a frozen-CLIP linear probe. On an independent dataset with non-overlapping labels, it also outperforms vanilla CLIP when both use the same labelled target gallery and prototype construction. These results suggest that clinical language and disease taxonomy provide complementary supervision for fine-grained visual representation learning and flexible hierarchical classification.

Keywords: Hierarchical skin disease classification · Taxonomy-guided contrastive learning · Language supervision · Vision–language models · Prototype classification

1 Introduction

Skin disease recognition is naturally hierarchical. A broad pathological category may contain several related diagnoses that share appearance and clinical characteristics, yet ultimately require fine-grained distinction. Hierarchy-aware

⋆ Corresponding author.

learning therefore seeks to coordinate coarse and fine predictions through semantic distances, multiple label levels, or constrained decision structures [2,3,23]. For dermatology, however, the challenge is not merely to respect the label tree at prediction time: the representation itself should capture what related diseases share without suppressing the morphological evidence that separates their subclasses.

Contrastive representation learning provides a natural mechanism for shaping such geometry, but its behavior depends critically on how positive and negative relationships are defined [7,17,15]. Under parent-level supervision, samples from different subclasses of the same parent become positives. This encourages disease-family alignment but may also pull clinically distinct sibling diagnoses together. We refer to this coarse–fine conflict as *taxonomy collapse* in the parent-supervised setting. In our experiments, parent-level supervised contrastive learning raises Main-Class (MC) macro-F1 to 50.85, whereas Sub-Class (SC) macro-F1 remains only 14.05. Thus, a representation can become substantially better at recognizing the broad category without acquiring the discrimination required within it.

Language provides a complementary source of structure. Concept-based models make clinically meaningful attributes explicit [9,4], while medical vision–language models align images with reports, literature, or concept descriptions [18,8,13]. We take a related but distinct view: language can act as supervision even when the ultimate task remains visual classification. Existing labels and structured descriptors—including morphology, anatomical location, and symptoms—can be verbalized into multiple clinically meaningful descriptions that shape the visual representation itself. In this formulation, language specifies *what* semantic evidence should be retained, while the disease taxonomy identifies *which* related diagnoses require stronger separation.

We realize this idea through a taxonomy-guided vision–language framework combining multi-positive concept supervision with hierarchy-informed pair weighting. Same-subclass samples form positives, while same-parent/different-subclass pairs receive increased negative importance as taxonomically related sibling diagnoses. Unlike hard-negative selection based on current embedding similarity, this relation is supplied directly by the known disease hierarchy. Clinical concepts and taxonomy therefore provide complementary signals for organizing the shared image–text space.

The learned representation is subsequently used without a fixed disease classifier head. Gallery captions are grouped by class to form textual prototypes; MC prediction is followed by SC prediction restricted to the children of the predicted parent, while the nearest caption provides a clinically grounded exemplar description. This formulation also decouples the encoders from the output vocabulary: a new disease set can be introduced by replacing the gallery while keeping both encoders frozen.

Taxonomy-guided optimization raises SC macro-F1 from 14.05 to 32.01, while hierarchy-constrained inference and prototype fusion raise the complete framework to 54.21. A frozen-CLIP linear probe remains substantially below the complete framework, and on an independent dataset our representation also

outperforms vanilla CLIP when both use the same labelled target gallery and prototype construction. Accordingly, our contributions are:

1. a taxonomy-guided image–text objective that strengthens discrimination among clinically related sibling diagnoses;
2. multi-positive clinical language supervision that shapes the visual representation without requiring an explicit concept bottleneck at inference; and
3. gallery-based hierarchical inference supporting clinically grounded retrieval and retraining-free adaptation to a disjoint target vocabulary.

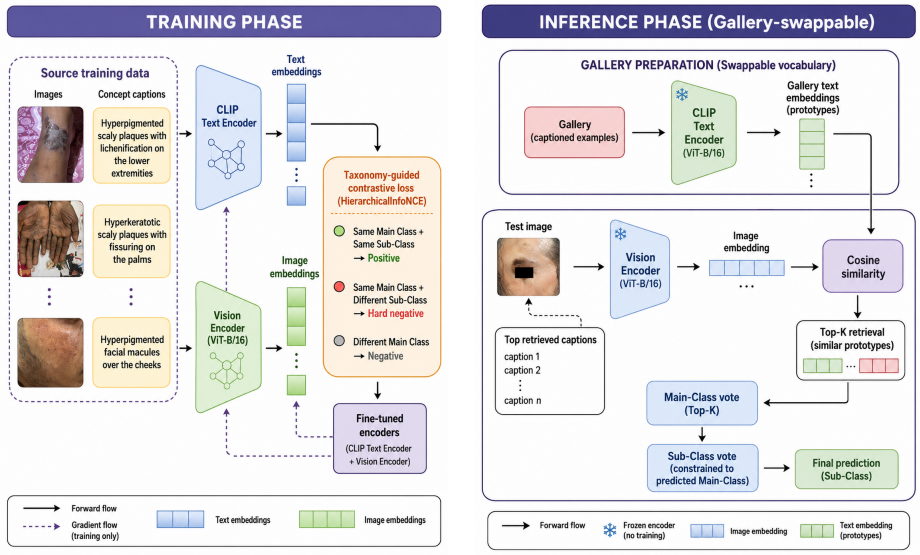
2 Related Work

Our setting intersects hierarchical classification, concept-based learning, medical vision–language representation learning, and prototype inference.

Hierarchical representation learning. Hierarchy-aware methods coordinate coarse and fine predictions through semantic distances, sequential decisions, or multiple label levels [2,3]. Barata *et al.* [1] condition successive skin-lesion decisions along a medical taxonomy, while Zhang *et al.* [23] incorporate multiple hierarchy levels into contrastive learning. HAC-LF [5], LightDPH [6], and HGCLIP [19] further explore hierarchy-aware skin or vision–language classification. DermaCon-IN [10] also evaluates hierarchical Concept Bottleneck Models (CBMs), where lesion and anatomical concepts form an explicit intermediate representation. Our approach instead uses structured concepts as language supervision of the shared image–text space, while the taxonomy directly determines which sibling classes receive increased negative importance.

Concept and language supervision. CBMs [9] and Concept Embedding Models [4] expose interpretable concepts within the prediction pathway. We instead express such information through multiple textual descriptions that supervise the representation without becoming mandatory intermediate predictions. Medical VLMs provide complementary domain-specific supervision: MedCLIP [18] incorporates medical semantics into image–text learning, BioMedCLIP [22] learns biomedical representations from large-scale scientific image–text data, MONET [8] learns dermatology representations grounded in medical literature, and Patrício *et al.* [13] use concept descriptions for skin-lesion diagnosis. Our framework couples concept-derived language supervision with explicit taxonomy-guided sibling discrimination.

Hard negatives and prototype inference. Hard-negative methods identify informative pairs through embedding similarity, sampling, or label relationships [17,15,16]. Our weighting instead follows the known taxonomy: every same-parent/different-subclass pair receives increased negative importance independently of its current embedding distance. For inference, CLIP [14] uses textual class prototypes, Tip-Adapter [21] adds a few-shot cache, and CHiLS [11] exploits hierarchical label



(a) Training phase **(b) Gallery-swappable inference**

Fig. 1. Overview of the proposed taxonomy-guided vision–language framework. (a) Concept-derived captions provide multi-positive language supervision, while same-Main-Class/different-Sub-Class pairs receive increased negative importance through HierarchicalInfoNCE. **(b)** Frozen encoders compare a test image with caption-derived prototypes for hierarchy-constrained prediction and caption retrieval. Replacing the gallery introduces a new disease vocabulary without encoder retraining.

sets. Our gallery operates on a representation already shaped by clinical concepts and taxonomy, supports constrained two-level prediction, and permits label-vocabulary replacement while keeping the encoders frozen.

Positioning this work. Rather than using language only as a multimodal input, output, or explanation, we use concept-derived language to supervise a visual representation, while the disease taxonomy specifies the sibling distinctions requiring additional separation. This combination links representation learning with deployment: the resulting space supports hierarchical classification, clinically grounded retrieval, and gallery-based adaptation to new disease vocabularies without encoder retraining.

3 Method

Figure 1 summarizes the framework. Source-domain training jointly shapes image and text representations using clinical language and disease-taxonomy supervision. At inference, both encoders remain frozen and the caption gallery defines the available prediction vocabulary.

3.1 Problem Setting and Concept Supervision

Let $\mathcal{D}_s = \{(x_i, m_i, s_i, \mathcal{C}_i)\}_{i=1}^{N_s}$, where x_i is an image, $m_i \in \mathcal{M}$ its Main Class (MC), $s_i \in \mathcal{S}(m_i)$ its Sub-Class (SC), and $\mathcal{C}_i = \{c_{i,r}\}_{r=1}^{R_i}$ its clinical captions. The source taxonomy contains 8 MCs and 19 SCs. Image encoder f_v and text encoder f_t produce normalized embeddings

$$\mathbf{z}_i^v = \mathrm{norm}(f_v(x_i)), \qquad \mathbf{z}_{i,r}^t = \mathrm{norm}(f_t(c_{i,r})), \qquad \mathrm{norm}(\mathbf{u}) = \frac{\mathbf{u}}{\|\mathbf{u}\|_2}. \quad (1)$$

CLIP's modality-specific projection heads map both outputs to the same d-dimensional space, enabling direct image–text similarity computation.

Source captions are generated from existing structured annotations, including disease label, lesion descriptors, anatomical location, symptoms, and available demographic fields. Qwen2 only verbalizes these supplied attributes into multiple descriptions; it does not receive the image or infer unannotated age, sex, location, or clinical findings. Thus, language provides alternative semantic views of existing annotations rather than new clinical labels. The paraphrases were not independently clinician-validated and are used only as representation-level augmentation, with their clinical content anchored to the underlying annotations. Unlike an explicit concept bottleneck, these concepts supervise the shared representation directly and are not required as intermediate predictions at inference.

For the target dataset, which lacks free-text descriptions, captions are constructed deterministically as `Skin lesion - [label]: Age: [a]. Location: [r]. Features: [f].` Although source and target disease labels are disjoint, their captions share overlapping clinical vocabulary through anatomical and symptom descriptors.

3.2 Taxonomy-Guided HierarchicalInfoNCE

With MC-level supervised contrastive learning, samples from different SCs under the same parent become positives. This favors coarse alignment but may suppress distinctions required for fine-grained diagnosis, a coarse–fine conflict we term *taxonomy collapse* in the parent-supervised setting. We instead use the hierarchy to define three complementary pair relationships:

Taxonomic relation	Training role
Same SC: $s_i = s_j$	Positive
Sibling SCs: $m_i = m_j,\ s_i \neq s_j$	Weighted negative
Different MCs: $m_i \neq m_j$	Ordinary negative

$$\mathcal{P}_i = \{(j, r) \mid s_j = s_i\},$$
$$w_{ij} = \begin{cases} \alpha, & m_i = m_j,\ s_i \neq s_j, \\ 1, & \text{otherwise.} \end{cases} \quad (2)$$

Here $\mathcal{P}_i$ contains the positive captions for anchor i, while w_{ij} increases the contribution of sibling-class negatives; we use $\alpha = 2.0$. Hardness is therefore supplied by the known taxonomy rather than inferred from current embedding distance. This is a structural priority, not an assumption that every cross-parent pair is visually easy: such pairs remain negatives and may still be close in the learned space.

For a mini-batch caption set $\mathcal{T}_{\mathcal{B}}$, the image-to-text objective is

$$\begin{aligned} \ell_i^{v\to t} &= -\frac{1}{|\mathcal{P}_i|}\sum_{(j,r)\in\mathcal{P}_i}\log\frac{\exp\left(\mathbf{z}_i^v\cdot\mathbf{z}_{j,r}^t/\tau\right)}{\sum_{(k,q)\in\mathcal{T}_{\mathcal{B}}} w_{ik}\exp\left(\mathbf{z}_i^v\cdot\mathbf{z}_{k,q}^t/\tau\right)}, \\ \mathcal{L}_{\text{hier}} &= \tfrac{1}{2}\left(\mathcal{L}_{v\to t}+\mathcal{L}_{t\to v}\right). \end{aligned} \tag{3}$$

Here τ is learnable and initialized to 0.07; $\mathcal{L}_{v\to t}$ averages $\ell_i^{v\to t}$ over image anchors, and $\mathcal{L}_{t\to v}$ is defined analogously with text anchors. Weighting a sibling negative by α is equivalent to adding $\log\alpha$ to its denominator logit; the embedding itself is unchanged. The objective therefore strengthens sibling-class discrimination without imposing a complete metric ordering over the taxonomy.

3.3 Gallery-Based Hierarchical Inference

Unlike a fixed softmax classifier, the trained encoders output embeddings rather than disease logits. After training, gallery captions are grouped by their known class labels and averaged to form textual class prototypes; similarities between a test-image embedding and these prototypes act as class scores.

Let $\mathcal{G}_c$ denote the gallery captions associated with class c. For test image x, prototype construction and hierarchical prediction are

$$\begin{aligned} \mathbf{p}_c &= \text{norm}\left(\frac{1}{|\mathcal{G}_c|}\sum_{j\in\mathcal{G}_c}\text{norm}(f_t(c_j))\right), \\ \hat{m} &= \arg\max_{c\in\mathcal{M}}\mathbf{z}^v\cdot\mathbf{p}_c^{\text{MC}}, \\ \hat{s} &= \arg\max_{c\in\mathcal{S}(\hat{m})}\mathbf{z}^v\cdot\mathbf{p}_c^{\text{SC}}, \qquad \mathbf{z}^v=\text{norm}(f_v(x)). \end{aligned} \tag{4}$$

Restricting the SC search to children of $\hat{m}$ prevents cross-hierarchy assignments. The most similar gallery caption, $c^* = \arg\max_{c_j\in\mathcal{G}}\mathbf{z}^v\cdot\text{norm}(f_t(c_j))$, is additionally returned as a clinically grounded exemplar description, rather than a guaranteed causal explanation.

For target-domain evaluation, the source gallery is replaced by captions from the labelled target training partition; target test samples are never used to construct prototypes. Only caption embeddings and class prototypes are recomputed: the image encoder, text encoder, temperature, and all learned parameters remain fixed. This enables encoder-retraining-free prediction over a disease vocabulary disjoint from the source taxonomy, without source-to-target label mapping or architectural modification.

4 Experiments

4.1 Experimental Setup

Datasets. DermaCon-IN [10] is the source benchmark, comprising 5,450 clinical images from South Indian outpatient clinics under an 8-Main-Class (MC), 19-Sub-Class (SC) taxonomy. We use its subject-wise split and construct four

Table 1. Source-domain results (%). **(a)** MC comparison; † denotes a fixed-output supervised classifier and ‡ an independently reported DermaCon-IN result. **(b)** Cumulative MC/SC macro-F1 ablation.

(a) Main-Class comparison

Method	Acc.	Prec.	F1	AUC
Swin-B†	70.41	**69.83**	**69.69**	78.51
Hier. CBM-2†‡	69.90	68.82	69.08	77.01
CLIP-ZS	25.78	24.79	16.25	71.42
CLIP linear†	58.61	30.82	29.40	79.08
Ours	**71.04**	62.67	63.95	**80.78**

(b) Cumulative ablation

Component	MC F1	SC F1
CLIP-ZS	16.25	10.52
MC-SupCon	50.85	14.05
+ H-InfoNCE	51.95	32.01
+ Constr. decode	58.95	43.38
+ Prototype fusion	**63.95**	**54.21**

training captions per image from the available structured lesion and anatomical annotations. PAD-UFES-20 [12] provides the target benchmark: 2,298 smartphone images from 1,373 patients across six diagnoses. We use an 80/20 patient-level split (1,837/461) and construct target captions from age, anatomical region, Fitzpatrick type, and symptom metadata. The transfer setting therefore combines population, acquisition, metadata, and label-vocabulary shifts, with no overlap between source and target disease labels.

Implementation and controls. We fine-tune `openai/clip-vit-base-patch16` (150M parameters) for 10 epochs using AdamW, batch size 32, cosine annealing, learning rate 2×10^{-5}, weight decay 10^{-4}, and $0.1\times$ learning rate for the vision encoder. Temperature τ is learnable and initialized to 0.07; training uses one NVIDIA A100. At target evaluation, all learned parameters remain frozen and only gallery embeddings and prototypes are recomputed. Following reviewer requests, we additionally evaluate two controls: (i) independent MC and SC linear probes trained on frozen vanilla-CLIP image embeddings, and (ii) vanilla CLIP using exactly the same labelled target training captions, patient split, and prototype construction as our method. These controls separate the learned-representation contribution from that of a downstream classifier and target-gallery supervision.

4.2 Source-Domain Results

Table 1(a) compares our framework with conventional, hierarchy-aware, and frozen-feature classifiers. The frozen-CLIP linear probe improves over CLIP-ZS to 29.40 MC F1 (20.21 SC F1), but remains well below our 63.95 MC F1 and 54.21 SC F1, indicating that a conventional classifier on generic CLIP features does not explain the gain. Our method achieves the highest MC accuracy and AUC, while Swin-B and the hierarchical CBM retain higher MC F1. The independently reported Type-2 hierarchical CBM also reaches 54.49 SC F1 on DermaCon-IN, close to our 54.21; unlike its explicit concept bottleneck and fixed classification heads, our framework uses structured concepts as language supervision and retains gallery-based inference.

Table 1(b) isolates the contributions of the proposed pipeline. MC-SupCon substantially improves coarse recognition but leaves SC F1 at 14.05. Hierarchical-InfoNCE raises SC F1 to 32.01, while constrained decoding and prototype fusion

Table 2. Transfer to the disjoint six-class PAD-UFES-20 vocabulary (%). † denotes target-trained contextual references. The two gallery-based methods use identical labelled target-training captions and prototype construction while keeping their encoders frozen.

Method	Encoder training	Target adaptation	F1	Bal. Acc.
ResNet-50† [12]	Target-trained	Full target training	67.8	58.3
DiffMIC† [20]	Target-trained	Full target training	67.1	64.6
Swin-B	Source-trained	Fixed source head	0.0	≈14.3
CLIP-ZS	Pretrained	Class-name prompts	18.40	22.70
CLIP + matched gallery	Pretrained	Matched target gallery	12.13	23.86
Ours	Source-trained	Gallery replacement	**24.38**	**31.33**

further increase it to 43.38 and 54.21, respectively. The gains therefore reflect complementary contributions from taxonomy-guided representation learning and hierarchy-aware inference.

4.3 Transfer through Gallery Replacement

Table 2 evaluates adaptation to a disjoint label vocabulary. The target-trained ResNet-50 and DiffMIC are contextual references rather than direct transfer baselines, while the fixed-output Swin-B cannot naturally predict over the new vocabulary. CLIP-ZS uses only class-name prompts and therefore receives less target supervision than our gallery-based method.

The matched-gallery control removes this confound by giving vanilla CLIP exactly the same labelled target-training captions, patient split, and prototype construction as our framework. Under this identical gallery, vanilla CLIP reaches 12.13 macro-F1 and 23.86 balanced accuracy, whereas our source-trained representation reaches 24.38 and 31.33, respectively. Thus, the transfer advantage persists when target-gallery supervision is held constant, supporting the contribution of the taxonomy- and language-shaped representation rather than the gallery alone.

Gallery replacement requires no gradient updates, architectural modification, or source-to-target label mapping. Because the gallery is constructed from labelled target-training metadata, this setting constitutes frozen-encoder gallery adaptation rather than label-free zero-shot transfer; target test samples are never used in prototype construction.

5 Conclusion

We presented a taxonomy-guided vision–language framework for hierarchical skin disease classification that brings clinical language and disease structure into representation learning itself. Multi-positive captions derived from existing labels and structured descriptors provide clinically grounded language supervision, while HierarchicalInfoNCE gives greater negative importance to same-parent/different-subclass pairs requiring fine-grained separation. This raises SC macro-F1 from

14.05 to 32.01, with the complete hierarchical inference framework reaching 54.21. A frozen-CLIP linear probe remains substantially below the complete framework, and on a disjoint target vocabulary our representation also outperforms vanilla CLIP when both use the same labelled gallery and prototype construction. These controls indicate that the gains arise from the learned representation rather than from a conventional downstream classifier or access to target-gallery supervision alone.

More broadly, our results suggest that language need not be useful only when a system ultimately consumes or generates text. It can instead act as structured supervision for a fundamentally visual task: clinical concepts indicate the evidence that the representation should retain, while taxonomy identifies the related diagnoses that require stronger separation. This supervision is built from existing annotations rather than requiring additional expert-written reports or mandatory concept predictions at deployment, while the resulting prototype space permits encoder-retraining-free adaptation to new disease vocabularies. Although the present evaluation is limited to two datasets and a single vision–language backbone, the caption construction and taxonomy-guided objective are not inherently tied to a particular dataset or architecture. Future work will study their generality across medical vision–language backbones and other skin-image tasks, including retrieval, segmentation, severity assessment, and out-of-distribution recognition.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Barata, C., Celebi, M.E., Marques, J.S.: Explainable skin lesion diagnosis using taxonomies. Pattern Recognition **110**, 107413 (2021). https://doi.org/https://doi.org/10.1016/j.patcog.2020.107413, https://www.sciencedirect.com/science/article/pii/S0031320320302168 3
2. Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., Lord, N.A.: Making better mistakes: Leveraging class hierarchies with deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12506–12515 (2020) 2, 3
3. Chang, D., Pang, K., Zheng, Y., Ma, Z., Song, Y.Z., Guo, J.: Your "flamingo" is my "bird": Fine-grained, or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11476–11485 (2021) 2, 3
4. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lió, P., Jamnik, M.: Concept embedding models: Beyond the accuracy–explainability trade-off. In: Advances in Neural Information Processing Systems. vol. 35, pp. 21400–21413 (2022) 2, 3
5. Hsu, C.Y., Chang, C.J., Lu, H.H.S.: Hac-lf: Hierarchy-aware contrastive learning framework for skin lesion classification. Preprint (2022) 3
6. Hsu, C.Y., Lu, H.H.S.: Lightdph: Lightweight dual-path hierarchical network for fine-grained skin lesion classification. Preprint (2024) 3

7. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 18661–18673 (2020) 2
8. Kim, C., Gadgil, S.U., DeGrave, A.J., Omiye, J.A., Cai, Z.R., Daneshjou, R., Lee, S.I.: Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine **30**, 1154–1165 (2024). https://doi.org/10.1038/s41591-024-02887-x 2, 3
9. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: Proceedings of the 37th International Conference on Machine Learning (ICML). pp. 5338–5348 (2020) 2, 3
10. Madarkar, S.S., Madarkar, M., Venkatesh, M., Prakash, T., Mopuri, K.R., MV, V., Sathwika, K., Kasturi, A., Raj, G., Supranitha, P., Udai, H.: DermaCon-IN: A multiconcept-annotated dermatological image dataset of indian skin disorders for clinical ai research. In: Advances in Neural Information Processing Systems. vol. 38 (2025). https://doi.org/10.52202/085713-3742, datasets and Benchmarks Track 3, 6
11. Novack, Z., McAuley, J., Lipton, Z.C., Garg, S.: CHiLS: Zero-shot image classification with hierarchical label sets. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 26342–26362. PMLR (2023) 3
12. Pacheco, A.G.C., Lima, G.R., Salomao, 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 Espirito Santo, R.B., Macedo, T.L.S.F., 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 in Brief **32**, 106221 (2020) 7, 8
13. Patrício, C., Teixeira, L.F., Neves, J.C.: Towards concept-based interpretability of skin lesion diagnosis using vision-language models. In: 2024 IEEE International Symposium on Biomedical Imaging. pp. 1–5. IEEE (2024). https://doi.org/10.1109/ISBI56570.2024.10635623 2, 3
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 8748–8763 (2021) 3
15. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. In: International Conference on Learning Representations (ICLR) (2021) 2, 3
16. Suresh, V., Ong, D.C.: Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 4381–4394. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.359 3
17. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5022–5030 (2019) 2, 3
18. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3876–3887 (2022) 2, 3

19. Xia, P., Yu, X., Hu, M., Ju, L., Wang, Z., Duan, P., Ge, Z.: HGCLIP: Exploring vision-language models with graph representations for hierarchical understanding. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 269–280. Association for Computational Linguistics, Abu Dhabi, UAE (2025) 3
20. Yang, Y., Fu, H., Aviles-Rivero, A.I., Schönlieb, C.B., Zhu, L.: Diffmic: Dual-guidance diffusion network for medical image classification. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 95–105 (2023) 8
21. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 493–510 (2022) 3
22. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image–text pairs (2023) 3
23. Zhang, S., Xu, R., Xiong, C., Ramaiah, C.: Use all the labels: A hierarchical multi-label contrastive learning framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16660–16669 (2022) 2, 3