

TOWARDS CONCEPT-BASED INTERPRETABILITY OF SKIN LESION DIAGNOSIS USING VISION-LANGUAGE MODELS

Cristiano Patrício^{1,3}, Luis F. Teixeira^{2,3}, João C. Neves¹

¹Universidade da Beira Interior and NOVA LINCS,

²Faculdade de Engenharia da Universidade do Porto, ³INESC TEC

ABSTRACT

Concept-based models naturally lend themselves to the development of inherently interpretable skin lesion diagnosis, as medical experts make decisions based on a set of visual patterns of the lesion. Nevertheless, the development of these models depends on the existence of concept-annotated datasets, whose availability is scarce due to the specialized knowledge and expertise required in the annotation process. In this work, we show that vision-language models can be used to alleviate the dependence on a large number of concept-annotated samples. In particular, we propose an embedding learning strategy to adapt CLIP to the downstream task of skin lesion classification using concept-based descriptions as textual embeddings. Our experiments reveal that vision-language models not only attain better accuracy when using concepts as textual embeddings, but also require a smaller number of concept-annotated samples to attain comparable performance to approaches specifically devised for automatic concept generation.

Index Terms— Concept-based Models, Interpretability, Skin Cancer, Vision-Language Models, Dermoscopy

1. INTRODUCTION

Automated computer-aided diagnosis systems for disease detection from medical images have undergone a remarkable increase in performance, primarily attributed to the enhanced capabilities of deep learning models. This paradigm has led to a substantial increase in the precision of these systems in providing accurate diagnosis in various medical image tasks, such as skin lesion diagnosis, assuring in some cases results that match the performance of dermatologists [1, 2]. However, the “black-box” nature of these deep learning-based systems in dermatology poses the most significant barrier to their broad adoption and integration into clinical workflow [3]. To alleviate this problem, interpretability methods have emerged to ensure the transparency and robustness of medical AI systems. Among these interpretable strategies, Concept Bottleneck Models (CBM) [4] are growing in popularity in medical imaging analysis [5, 6, 7], since they allow to explain the decision process based on the presence or absence of human-

understandable concepts, which aligns perfectly with the way clinicians draw conclusion from medical images. Furthermore, several studies concluded that humans prefer concept-based explanations over other forms of explanations, such as heatmaps or example-based [8]. In spite of their popularity, the development of concept-based models depends on dense annotations of human-understandable concepts [9], which are time-consuming and require expertise from domain experts, limiting the adoption of such models in medical image tasks. Several works [9, 10, 11, 12] attempt to mitigate this problem by querying Large Language Models (LLMs) to generate additional information about target classes to form candidate concepts.

In this work, we show that despite these advances, detailed concept-based descriptions generated from LLMs lead to inferior classification accuracy when compared with the use of textual embeddings derived directly from dermatoscopic concepts. Specifically, we compared the performance of LLMs on three well-known skin lesion datasets [13, 14, 15] using three distinct strategies for measuring the similarity between a given query skin image and textual embeddings: (i) utilizing the target class as textual embedding; (ii) using a set of dermoscopic concepts annotated by board-certified dermatologists as textual embeddings; and (iii) leveraging concept descriptions generated by ChatGPT as textual embeddings. Our experiments reveal that (i) relying on expert-selected dermoscopic concepts as textual embeddings leads to better performance in distinguishing melanoma from other diseases, in addition to providing concept-based explanations and (ii) a simple and efficient embedding learning procedure on top of feature embeddings of CLIP [16] could attain comparable performance to models specifically designed for the task of automated concept generation of dermoscopic features.

Our contributions can be summarized as follows: (i) we introduce an efficient and simple embedding learning procedure to improve the performance of CLIP models in the downstream task of melanoma diagnosis; (ii) we alleviate the annotation burden of CBMs by using zero-shot capabilities of Vision Language Models (VLMs) to automatically annotate concepts; (iii) we provide concept-based explanations for the model prediction based on expert-selected dermoscopic concepts.

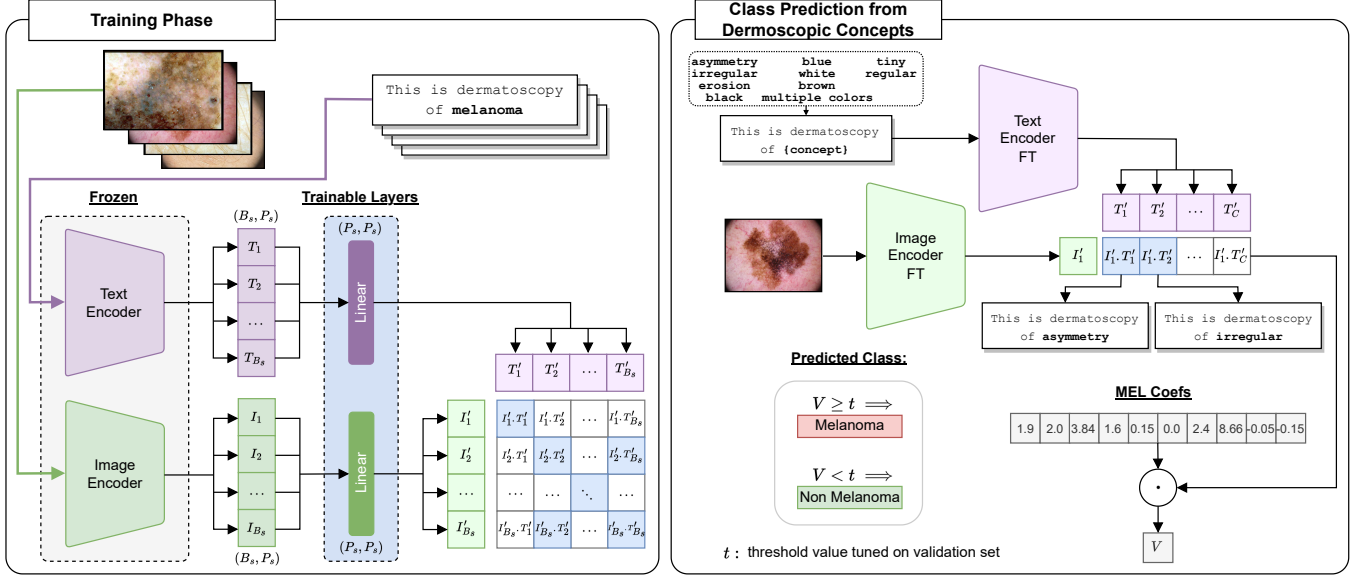


Fig. 1: The workflow of our proposed strategy. After learning the new multi-modal embedding space (left), we predict the presence of melanoma by linearly combining the similarity scores with the melanoma coefficients acting as the bottleneck layer of CBM. The result of this operation is then compared with a threshold value to predict the presence or absence of melanoma.

2. METHOD

Figure 1 presents an overview of the proposed method. The training phase consists of learning a new multi-modal embedding space for approximating image and textual embeddings of the same category (section 2.1). The learned projection layers are then used to calculate the feature embeddings of both the image and textual descriptions in order to predict melanoma (section 2.2) by: (i) calculating the cosine similarity between the image feature and the text encoding of each disease label; (ii) calculating the cosine similarity between the image and a concept c in the concept set C , whose scores are then fed into the classification layer to determine the presence of melanoma; or (iii) calculating the cosine similarity between the image and a set of m concept descriptors per concept c , average the scores per concept, and then fed into the classification layer as in (ii).

2.1. Embedding Learning

Let $\mathcal{D} = \{(i, y)\}$ be a batch of image-label pairs where i is the image and $y \in \mathcal{Y}$, is a label from a set of N classes. We extract the features of the frozen CLIP image encoder $\mathcal{I}(\cdot)$ and the text encoder $\mathcal{T}(\cdot)$ to obtain the feature embedding of the image $x = \mathcal{I}(i) \in \mathbb{R}^d$ and the feature embedding of the label $l = \mathcal{T}(y) \in \mathbb{R}^d$. The training phase (Figure 1) thus consists of learning a new multi-modal embedding space by jointly training an image projection layer W_I and text projection layer W_T to maximize the cosine similarity of the image feature $W_I.x$ and text feature $W_T.l$ embeddings of the n pairs sharing the same disease while minimizing the cosine similarity

of embeddings of the pairs from different diseases. For this, we define a target matrix as having ones on image-label pairs sharing the same disease label, and zeros in the remaining pairs. We adopt the objective function used in [16].

2.2. Strategies for Melanoma Diagnosis

Baseline The most straightforward strategy for using CLIP in the task of melanoma classification is to calculate the similarity between the visual descriptor of the image $x = \mathcal{I}(i)$ and the textual feature representation of the N disease labels $l = \mathcal{T}(y)$, $y \in \mathcal{Y}$. The predicted disease label is given by $\hat{y} = \arg\max_{y \in \mathcal{Y}} S_c(W_I.x, W_T.l)$, where S_c is the cosine similarity.

CBM Alternatively, we can calculate the degree to which a dermoscopic concept $c \in C = \{c_1, \dots, c_{N_c}\}$ is present in the image by measuring the similarity between the feature embedding of the image $x = \mathcal{I}(i)$ and each feature embedding of concept c given by $E_C \in \mathbb{R}^{N_c \times d}$, where each row of E_C is a text feature $\mathcal{T}(c) \in \mathbb{R}^d$ of a concept c . Then, we employ the dermoscopic concept coefficients (**MEL Coefs** in Figure 1) extracted from a previously trained linear model for melanoma prediction [17], denoted as $\mathcal{W}_{mel} \in \mathbb{R}^{1 \times N_c}$, and multiply them with the obtained concept scores $p = S_c(W_I.x, W_T.E_C)$, $p \in \mathbb{R}^{N_c \times 1}$. Let $V = \mathcal{W}_{mel} \cdot p$. The final prediction is thus given by $\hat{y} = \begin{cases} 0, & \text{if } V < t \\ 1, & \text{if } V \geq t \end{cases}$ where t is a threshold value tuned on the validation set.

GPT + CBM We query ChatGPT with a designed prompt to generate a set of m textual descriptions for a given dermoscopic concept c . The chosen prompt ‘‘According to pub-

lished literature in dermatology, which phrases best describe a skin image containing `{concept}`?” returns a total of five descriptions for each individual concept c (see supplementary). We obtain the feature embedding for the m descriptions $E_{sc} = \mathcal{T}(s_1^c, \dots, s_m^c)$, $E_{sc} \in \mathbb{R}^{m \times d}$ of a concept c . We calculate the concept scores as $p_c = \frac{1}{m} \sum_{i=0}^m S_c(W_I \cdot x, W_T \cdot E_{s_i^c})$. Let $V = \mathcal{W}_{mel} \cdot \sum_{c=0}^{N_c} p_c$. The final score indicating the presence of melanoma is thus given by $\hat{y} = \begin{cases} 0, & \text{if } V < t \\ 1, & \text{if } V \geq t \end{cases}$.

3. EXPERIMENTAL SETUP

We evaluate different CLIP variations, using our proposed embedding learning, and compare it with MONET [17], a foundation model trained on dermatological images, under the previously defined strategies (section 2.2) on three dermoscopic datasets. Also, we report the performance of a black-box linear probing model to assess whether our approach can maintain black-box accuracy without compromising interpretability.

Datasets Three dermoscopic datasets were selected for our experiments, namely: PH² [13], Derm7pt [14] and ISIC 2018 [15]. The PH² dataset encompasses dermoscopic images of melanocytic lesions, including “melanoma” and two types of “nevus” that were merged and treated as singular “nevus”. For PH², we used 5-fold cross-validation. Derm7pt comprises clinical and dermoscopic images, which we filtered to obtain images of “nevus” and “melanoma” classes. ISIC 2018 is composed of dermoscopic images including different types of skin lesions, namely “melanoma”, “melanocytic nevus”, “basal cell carcinoma”, “actinic keratosis”, “benign keratosis”, “dermatofibroma”, and “vascular lesion”. Detailed statistics of the datasets, including the train/val/test splits, are presented in Table 1¹.

Dataset	Classes	Train size	Validation size	Test size
PH ² [13]	2	160 (28 to 34)	-	40 (6 to 12)
Derm7pt [14]	2	346 (90)	161 (61)	320 (101)
ISIC 2018 [15]	7	8,012 (890)	2,003 (223)	1,511 (171)

Table 1: Dataset statistics. Numbers between rounded brackets represent the # of Melanoma examples in the split.

3.1. Implementation Details

Embedding Learning The projection layers (section 2.1) were trained on Derm7pt and ISIC 2018 datasets using the AdamW optimizer with a learning rate of $1e^{-5}$. Also, a learning rate decrease policy was used with a patience of 1 and a factor of 0.8. The trainable projection layers are linear layers with the same dimension of the output of image and text

¹We followed the split partition adopted in [14] for the Derm7pt dataset and in [18] for ISIC 2018.

encoder of CLIP². For the evaluation of MONET we follow the proposed strategy by the authors to calculate the concept scores. For the black-box linear probing, we follow [16] and use image features taken from the penultimate layer of each model, ignoring any classification layer provided. A logistic regression classifier is trained on the top of the extracted image features using scikit-learn’s L-BFGS implementation, with maximum 1,000 iterations.

Preprocessing The input images were preprocessed according to the transformations defined in the original image encoders of CLIP variations. Additionally, and following [7], we use segmented versions of the images. This strategy ensures that solely the area of the lesion is considered, preventing the model from giving attention to artifacts in the image. Most importantly, this procedure allows improving the final classification results.

4. RESULTS

4.1. Quantitative Analysis

Comparison with Original CLIP and MONET Table 2 compares the performance of the original CLIP model with our method across three different strategies on two datasets. The reported results represent the average Balanced Accuracy (BACC) obtained across CLIP model variations for each specific strategy. Our method outperforms CLIP original variations by an average of 11.5% and 9.2% on both datasets, respectively. The most significant improvement is observed on the Baseline strategy for Derm7pt, and on CBM strategy for ISIC 2018. Figure 3 (left) shows the efficiency of our method in comparison to the MONET model. Notably, our method achieves a comparable level of performance of MONET while requiring significantly less training time. On the other hand, Figure 3 (right) depicts the evolution of AUC (in %) as more image-label pairs are added into the training set of ISIC 2018. The results show that CLIP RN50, CLIP ViT L/14 and CLIP ViT-B/32 attain comparable performance with MONET when using only between 40-60 image-label pairs in the training set.

Strategy	Derm7pt [14]		ISIC 2018 [15]	
	Orig.	Ours	Orig.	Ours
Baseline	61.3 ± 2.4	75.0 ± 2.5	54.1 ± 5.0	63.2 ± 1.4
CBM	65.4 ± 2.6	75.4 ± 2.3	60.6 ± 3.0	70.4 ± 3.0
GPT+CBM	64.1 ± 6.3	74.9 ± 2.6	61.2 ± 3.2	69.9 ± 3.2

Table 2: Performance gains of CLIP with our proposed embedding learning strategy in terms of BACC.

Evaluation of different VLMs for melanoma diagnosis The results presented in Figure 2 show the performance in terms of BACC. For the PH² dataset, the results represent the average performance over 5-fold cross-validation. The results

²The source code and supplementary material are available at <https://github.com/CristianoPatricio/concept-based-interpretability-VLM>

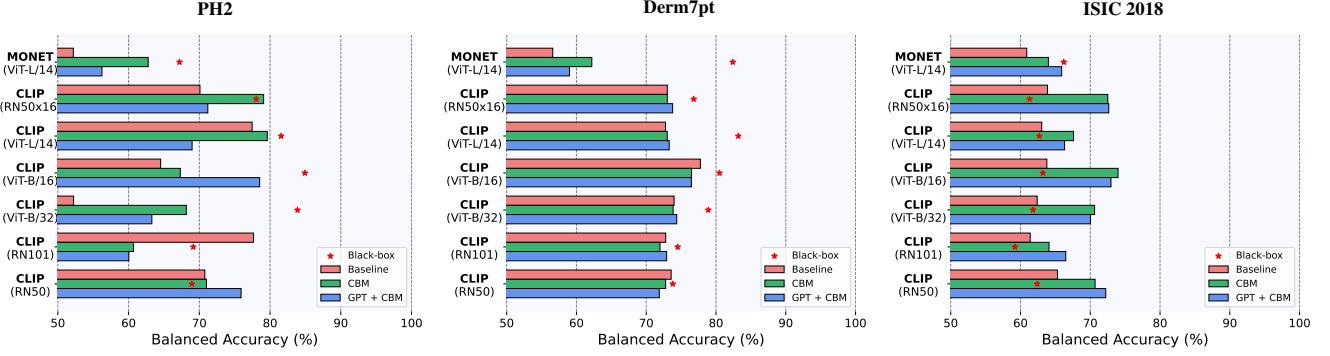


Fig. 2: Evaluation results (in BACC %) of the different classification strategies (Baseline, CBM and GPT+CBM) on three datasets (PH², Derm7pt and ISIC 2018) for melanoma detection. Black-box linear probing performance is marked with $\star$.

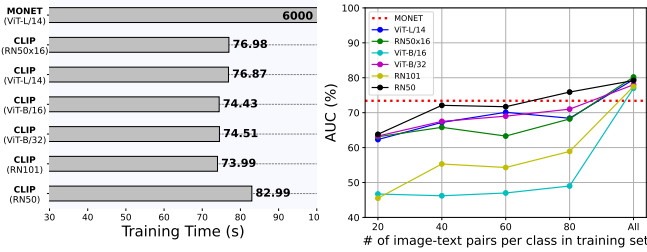


Fig. 3: Computational performance analysis of our proposed embedding learning procedure.

reported for Derm7pt and ISIC 2018 datasets are the averages obtained from four separate runs. The results on PH² dataset suggest that the GPT-CBM strategy outperforms both the Baseline and CBM strategies for CLIP ViT-B/16. Additionally, the CBM strategy demonstrates statistically significant improvement over the GPT+CBM strategy when applied to RN50x16. Regarding the Derm7pt dataset, all strategies exhibit comparable performance. However, a marginal gain of GPT+CBM over CBM and the Baseline is noticeable in 4 out of 7 models. In the case of ISIC 2018, the results show significant improvement of both CBM and GPT+CBM strategies over the Baseline ($p < 0.05$).

4.2. Interpretability by Dermoscopic Concepts

Utilizing dermoscopic concepts for melanoma detection ensures the interpretability and transparency of the model’s decision-making process. In Figure 4, we present two illustrative examples, each accompanied by the predicted dermoscopic concepts. In the upper image, the model classifies it as non-melanoma, as indicated by the negative contributions of dermoscopic concepts typically associated with melanoma. Conversely, the lower image was correctly classified as melanoma, as evidenced by the positive contributions of melanoma-specific concepts, which align with the ABCDEs of melanoma [19]. Additional examples can be

found in the supplementary material.

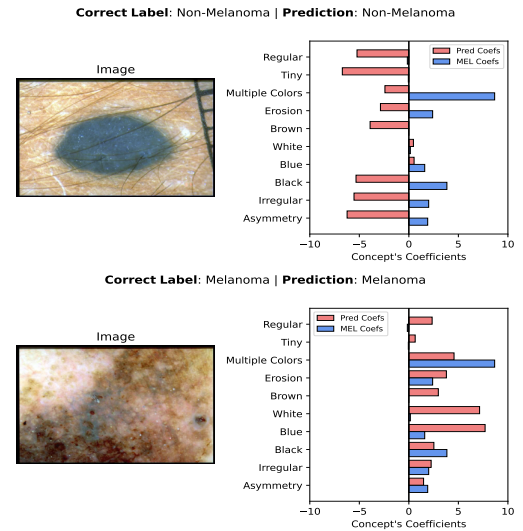


Fig. 4: Examples of dermoscopic images classified based on dermoscopic concepts.

5. CONCLUSIONS AND FUTURE WORK

This paper presents an efficient embedding learning procedure to enhance the performance of CLIP models in the downstream task of melanoma diagnosis, utilizing various strategies. Our comparative evaluation of VLMs’ efficacy in melanoma diagnosis indicates that predicting melanoma based on expert-selected dermoscopic concepts is more reliable than using the textual description of the target class, promoting interpretability in decision-making. Additionally, our experiments suggest that incorporating detailed descriptions of concepts as a proxy to use them directly in predicting melanoma does not lead to statistically significant improvements. In future research, we plan to expand the analysis to other imaging modalities to foster trust and acceptance of automated diagnosis systems in daily clinical practices.

Acknowledgments This work was funded by the Portuguese Foundation for Science and Technology (FCT) under the PhD grant “2022.11566.BD”, and supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT.IP.

Compliance with ethical standards This research study was conducted using human subject data, available in open access. Ethical approval was not required.

6. REFERENCES

- [1] Noel CF Codella, Q-B Nguyen, Sharath Pankanti, David A Gutman, Brian Helba, Allan C Halpern, and John R Smith, “Deep learning ensembles for melanoma recognition in dermoscopy images,” *IBM Journal of Research and Development*, vol. 61, no. 4/5, pp. 5–1, 2017.
- [2] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [3] Veronica Rotemberg, Allan Halpern, Steven Dusza, and Noel CF Codella, “The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice,” in *Seminars in Cutaneous Medicine and Surgery*, 2019, vol. 38, pp. E38–E42.
- [4] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang, “Concept Bottleneck Models,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 5338–5348.
- [5] Zhengqing Fang, Kun Kuang, Yuxiao Lin, Fei Wu, and Yu-Feng Yao, “Concept-based Explanation for Fine-grained Images and Its Application in Infectious Keratitis Classification,” in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 700–708.
- [6] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed, “On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors,” in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–10.
- [7] Cristiano Patrício, João C. Neves, and Luis F. Teixeira, “Coherent Concept-based Explanations in Medical Image and Its Application to Skin Lesion Diagnosis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3799–3808.
- [8] Vikram V Ramaswamy, Sunnie SY Kim, Ruth Fong, and Olga Russakovsky, “Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10932–10941.
- [9] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar, “Language in a bottle: Language model guided concept bottlenecks for interpretable image classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19187–19197.
- [10] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng, “Label-Free Concept Bottleneck Models,” *arXiv preprint arXiv:2304.06129*, 2023.
- [11] Sachit Menon and Carl Vondrick, “Visual classification via description from large language models,” *arXiv preprint arXiv:2210.07183*, 2022.
- [12] An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, et al., “Robust and Interpretable Medical Image Classifiers via Concept Bottleneck Models,” *arXiv preprint arXiv:2310.03182*, 2023.
- [13] Teresa Mendonça, Pedro M. Ferreira, Jorge S. Marques, André R. S. Marcal, and Jorge Rozeira, “PH2 - A Dermoscopic Image Database for Research and Benchmarking,” in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 5437–5440.
- [14] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh, “Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019.
- [15] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, 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.
- [16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning Transferable Visual Models from Natural Language Supervision,” in *ICML*, 2021, pp. 8748–8763.
- [17] Chanwoo Kim, Soham Uday Gadgil, Alex J DeGrave, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee, “Fostering transparent medical image AI via an image-text foundation model grounded in medical literature,” *medRxiv*, pp. 2023–06, 2023.
- [18] Catarina Barata, Veronica Rotemberg, Noel CF Codella, Philipp Tschandl, Christoph Rinner, Bengu Nisa Akay, Zoe Apalla, Giuseppe Argenziano, Allan Halpern, Aimilios Lallas, et al., “A reinforcement learning model for AI-based decision support in skin cancer,” *Nature Medicine*, pp. 1–6, 2023.
- [19] Darrell S Rigel, Robert J Friedman, Alfred W Kopf, and David Polsky, “ABCDE—an evolving concept in the early detection of melanoma,” *Archives of Dermatology*, vol. 141, no. 8, pp. 1032–1034, 2005.