Interpretable Skin Cancer Diagnosis with Contrastive Language-Image Pre-training

Abstract

Recent advances in machine learning and computer vision have significantly improved the performance of skin cancer diagnostic models. However, their lack of interpretability poses a challenge for clinical adoption, as physicians may find it difficult to trust a diagnosis made by a “black box” system. We propose a novel methodology for skin cancer diagnosis using Contrastive Language-Image Pre-training (CLIP), allowing physicians to provide a set of features in natural language and then determine the weight our model gave each feature in its diagnosis. This approach aims to bridge the communication gap between physicians and machine learning models. We show that the CLIP model is able to diagnose skin cancer in a zero-shot setting and provide insight into how each provided feature contributes to its diagnosis.

Introduction

According to the World Health Organization (WHO), skin cancer is one of the most common types of cancer worldwide, with 2-3 million instances being diagnosed every year [1]. Skin cancers are mainly divided into two types: melanoma and non-melanoma. Melanoma generally has the highest mortality rate, and early detection of melanoma before metastasis significantly increases the chances of successful treatment [2]. The use of computer vision to improve melanoma diagnosis has thus been studied extensively, with the goal of being able to visually determine if a skin lesion is likely to be malignant. Older methods relied on classical techniques such as constructing hand-crafted feature detection (shape, color, etc.) [3] and the k-nearest neighbors algorithm [4]. As of late, this task has been approached primarily with deep learning techniques. Specifically, Convolutional Neural Networks (CNNs) have achieved remarkable success. For example, Esteva et al. [5] trained a CNN model on over 120 thousand images and achieved dermatologist-level diagnostic performance.

A significant hurdle to the clinical deployment of computer vision-based diagnostic systems is that CNNs and other deep learning techniques are “black boxes,” meaning their decision-making processes are not easily understood or interpreted by human users. As such, although they can be highly accurate, their reasoning is not transparent, making it challenging for dermatologists to trust and adopt these systems. This is of particular concern in the medical field, where the consequences of a misdiagnosis can be grave, and understanding why a particular diagnosis was made is crucial.

Researchers have attempted to make diagnostic systems more interpretable. For example, Mahmud et al. [6] analyzed several pre-trained CNN models to determine which features the models were using to make their diagnosis. Similar approaches were used by Mridha et al. [7] and Alfi et al. [8]. We attempt an alternate approach: instead of determining the features used by an ML model, we allow a physician to provide a set of features in natural language, and then determine the weight our model gave each feature in its diagnosis. In this way, we aim to bridge the communication gap between the physicians and the ML model, enabling the former to understand and trust the model's decision-making process further. To match natural language descriptions with medical images, we utilize the CLIP model.

CLIP (Contrastive Language-Image Pretraining) is a model developed by OpenAI that has shown promising results in connecting natural language explanations with images [9] and in medical image analysis [10]. This model is trained to understand a broad range of visual and textual inputs, making it possible to link dermatologists' written descriptions with corresponding images of skin lesions. The CLIP model is particularly suited for this task because of its ability to understand and map complex, high-dimensional data from both language and images.

Our methodology involves inputting the clinical descriptors into our system and the corresponding image of the lesion. The CLIP model then maps these inputs into a joint-embedding space wherein similarity can be measured. The model assigns weights to each descriptor based on this analysis. This not only allows the physician to see which characteristics the model considers most critical but also helps them comprehend the model's decision-making process.

Related Work

Several recent studies have focused on improving the interpretability of machine-learning models in dermatology. Li et. Al [11] systematically corrupted different parts of the input image and observed the changes in the output prediction. This enabled them to localize the areas of the image that the model was focusing on, therefore giving insight into what features it was using to classify the skin lesion. They found that dermatologists agreed with patterns found in the model's prediction. Likewise, Mahmud et al. [6] studied the interpretability of CNNs using class activation mapping and found that the extracted features largely align with dermatology texts. Mridha et al. [7] proposed a multiple-instance learning approach, which provides feature-level interpretability. In a departure from previous techniques, Barata et al. [12] used an attention mechanism, checking attention weights to assess which regions were most influential in a prediction. In all of these approaches, interpretability is achieved by post hoc analysis after a model has made its predictions.

This study departs from the existing ones by letting a dermatologist provide natural-language explanations that the model uses to determine the weightage of each feature in its diagnosis. Specifically, the CLIP model is used to link these explanations with the corresponding images.

Methods

Model

CLIP (Contrastive Language-Image Pretraining) is a model developed by OpenAI with the ability to understand a broad range of visual and textual inputs and map them into a common embedding space where their semantic similarity can be measured. This shared embedding space allows for complex comparisons and interactions between textual and visual data, providing a representation that can capture subtle details in both modalities.

CLIP uses a vision model and a language model to embed images and text into a shared embedding space. The vision model can be based on ResNet [13] or Vision Transformer [14], while the language model is a transformer-based model like BERT [15]. Both models are trained together with a contrastive learning objective, where the dot product between matching image-text pairs is minimized.

Dataset

To test the zero-shot performance of CLIP in melanoma detection, we collect a set of 3 diagnosis criteria for melanoma (written in natural language) with a doctor's supervision. For every criterion, we use 5 differently phrased variations. We tested the model on a dataset of 1800 images of malignant and benign moles provided by the International Skin Imaging Collaboration [16]. To show that the model is able to follow criteria interpretably, we collected 30 images for each criterion that were particularly indicative of the corresponding characteristic (e.g., if one of the criteria was asymmetry, these would be notably asymmetrical moles).

Methodology

We evaluate the model on two criteria: overall classification accuracy and usage of the provided diagnostic criteria. For the overall classification task, we run an inference with the CLIP image model on images of moles, then compare (using cosine similarity) the image embedding to the embedding of every criterion and to the term "benign mole." We use majority voting to classify the mole as either malignant or benign; if most of the criteria embeddings are more similar to the image than the benign mole embedding, the mole is classified as malignant and vice versa.

To evaluate the model's interpretability, we use the constructed dataset of particularly indicative images. For each diagnostic criterion, we calculate the degree to which that criterion aligned with the image. We expect that images representative of a specific criteria will have a higher cosine similarity with that diagnostic criteria's textual embedding compared to other criteria.

Results

In the overall classification task, the pre-trained CLIP model achieves a zero-shot F1 score of 0.71. While this score is lower than a CNN trained on our dataset (EfficientNet [17] scores an F1 of 0.82), it is achieved purely

through a text description and zero-shot prompting, rather than extensive labeled training, demonstrating the capabilities of the CLIP model in this application.

Table 1. Precision, recall, and F1-score of CLIP in overall classification

	Precision	Recall	F1
Positive label	0.78	0.64	0.71
Negative label	0.65	0.78	0.71

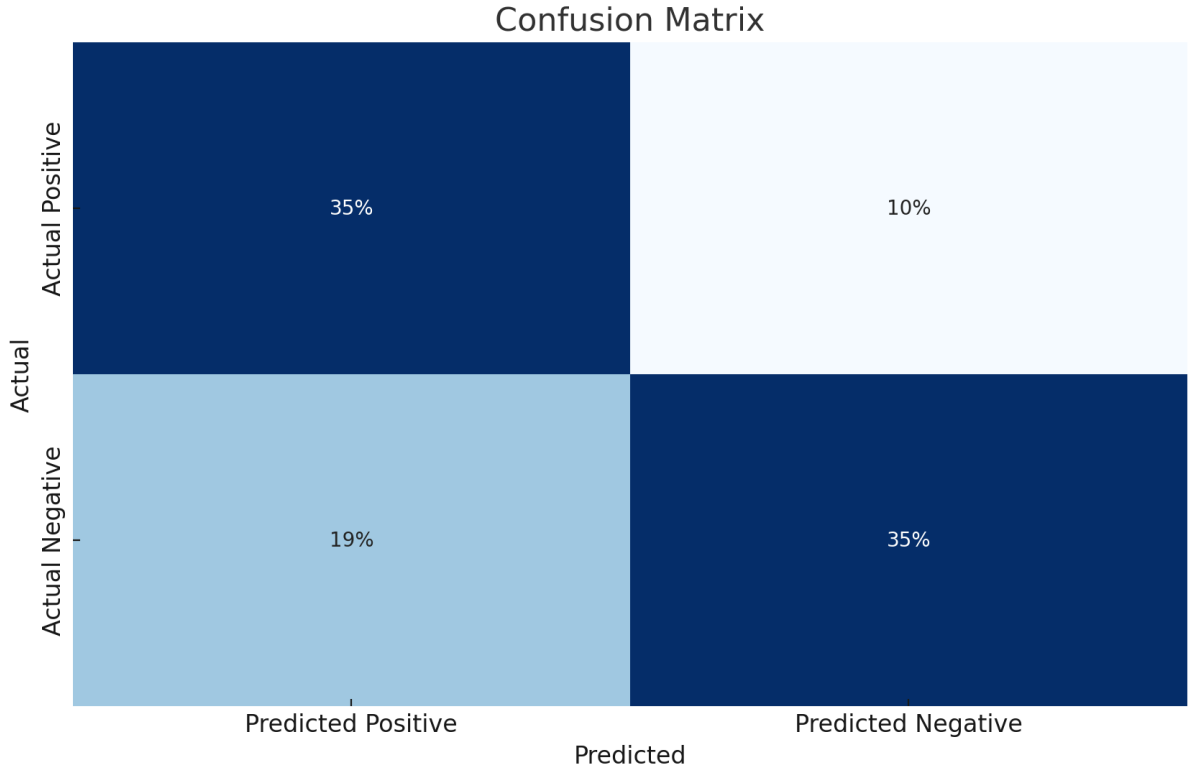

Figure 1. Confusion matrix of CLIP in overall classification

In the interpretability task, we find that the diagnostic criteria do indeed align as expected. Images representative of a certain criterion have a higher similarity score with the textual embeddings of that criterion. This indicates that if a mole image is particularly indicative of a criterion, such as 'asymmetry,' the model indeed gives more weight to 'asymmetry' in its diagnosis.

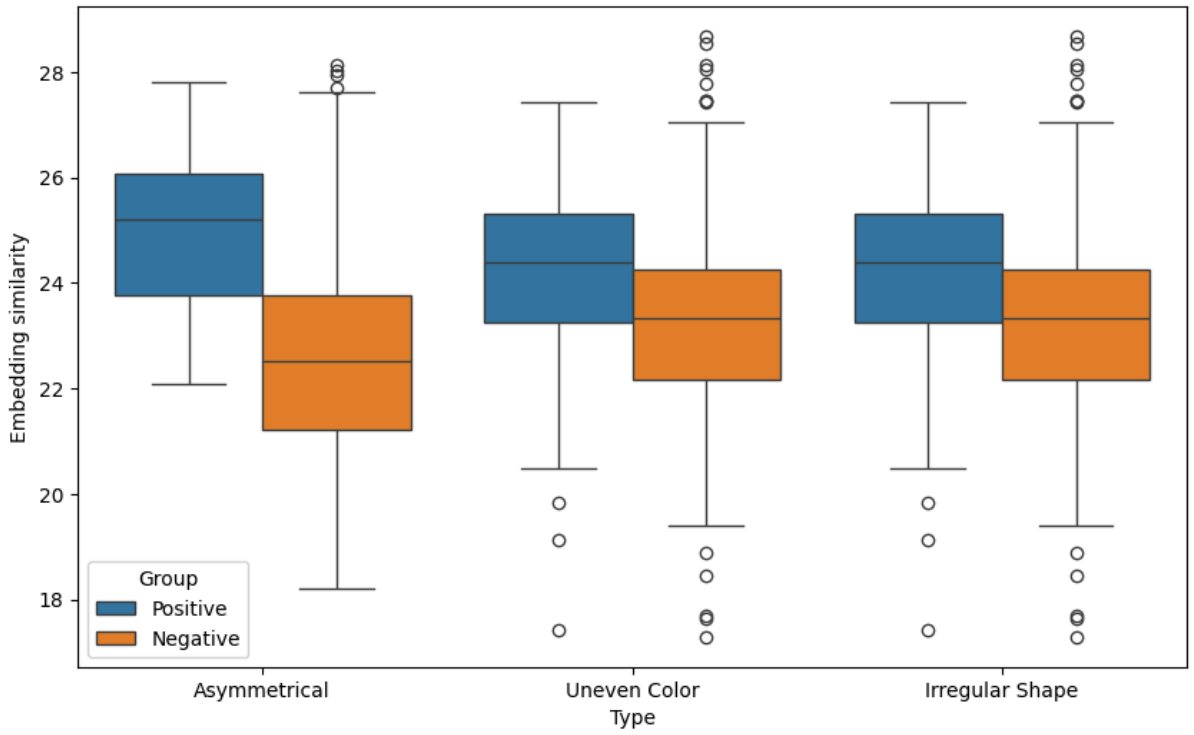

Figure 2. Computed similarities between a criterion and images particularly indicative of it (positive group) vs. a criterion and images not particularly indicative of it (negative group)

Error Analysis

When examining errors made by the model, the interpretability of CLIP allows us to gain insight into the cause of misclassification. In figure 3 we show an example of a type I error (false positive) and the corresponding similarity scores between the image and diagnosis criteria. We apply the softmax function to the similarity scores to create a probability-like weighting that shows the criteria that contributed the most when the model made its decision.

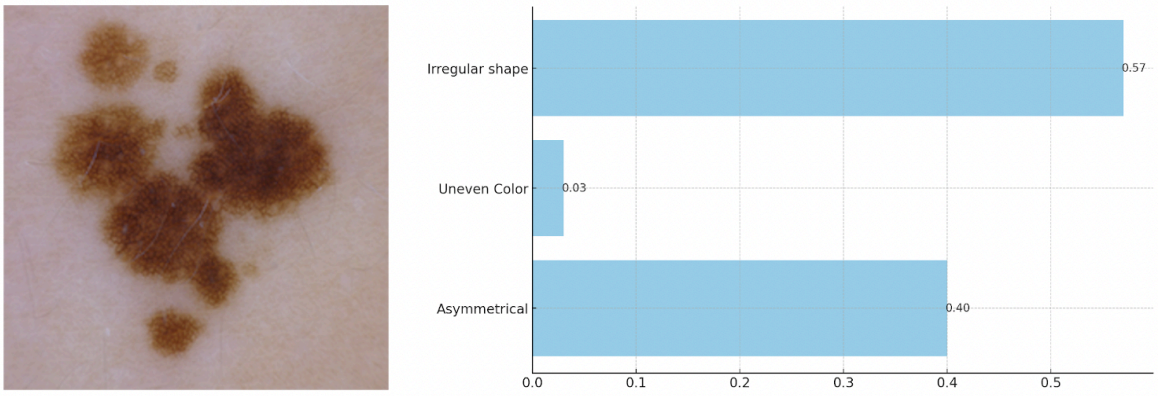

Figure 3. Image of a benign mole that was misclassified as malignant, and its corresponding similarities

The similarity score distribution in Figure 3 indicates that the moles' irregular shape and asymmetry were the primary reasons it was classified as malignant. This mole is indeed unusually uneven and asymmetric for a benign mole, hence the model's incorrect prediction. This highlights the strength of our approach: we are not only able to predict a diagnosis, but we can also understand why a particular diagnosis was made.

In Figure 4, we show an example of a type II error (false negative).

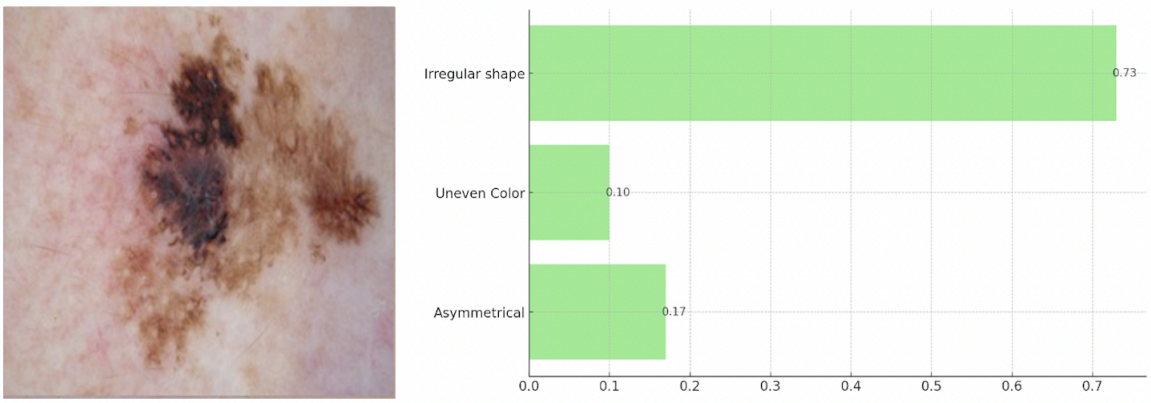

Figure 4. Image of a malignant mole that was misclassified as benign, and its corresponding similarities

The similarity score distribution in Figure 3 indicates that the moles' irregular shape was most concerning to the model. However, the model was not sufficiently concerned about it to classify this mole as malignant. This is presumably because the mole's other features were not indicative of melanoma according to the model. In this case, it seems the model failed to properly recognize the degree of uneven coloring and asymmetry in the mole.

Conclusion and Future Work

In this work, we demonstrate that pre-trained contrastive language-image models can bridge the gap between image-based skin lesion diagnosis and physicians' understanding by providing a suggestive explanation of the model's diagnostic process in the form of weights assigned to clinical descriptors. Our preliminary results suggest that the CLIP model can not only achieve a reasonable diagnostic accuracy, which further validates the applicability of the CLIP model in the medical imaging field, but it also holds the potential to improve the interpretability aspect of AI-based image diagnosis systems.

Our interpretability analysis was limited by the size of the criteria-specific datasets we were able to collect (due to the need for these datasets to be verified by a clinician). Future work should collect larger, more diverse datasets in order to more thoroughly validate this methodology. Additionally, future work should also explore the potential of fine-tuning the CLIP model on dermatology-specific datasets to improve its performance. We also suggest that novel visualization techniques should be explored for a more intuitive understanding of how descriptor weights contribute to the diagnostic decision.

References

1. https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer
2. Anthony F. Jerant, Jennifer T. Johnson, Catherine Demastes Sheridan, and Timothy J Caffrey. Early detection and treatment of skin cancer. American family physician, 62(2), 2000
3. M. Emre Celebi, Hassan A. Kingravi, Bakhtiyar Uddin, Hitoshi Iyatomi, Y. Alp Aslandogan, William V. Stoecker, and Randy H. Moss. A methodological approach to the classification of dermoscopy images. Computerized Medical Imaging and Graphics, 31(6):362–373, 2007.
4. https://ieeexplore.ieee.org/abstract/document/918473
5. [14] 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, 542(7639):115, 2017.
6. https://arxiv.org/pdf/2312.10696.pdf
7. https://ieeexplore.ieee.org/document/10107401

8. https://www.mdpi.com/2075-4418/12/3/726
9. https://arxiv.org/abs/2103.00020
10. https://arxiv.org/abs/2312.07353
11. https://ieeexplore.ieee.org/document/8857334
12. https://www.scopus.com/record/display.uri?eid=2-s2.0-85085040302&origin=inward&txGid=1effcfe746cdb46ca394b299b60bffdc
13. He et al., 2016
14. Dosovitskiy et al., 2020
15. Kenton and Toutanova, 2019
16. https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign