Explainable AI for Dermatology: Combining CNN Classification, Grad-CAM, and Vision–Language Models

Code: Colab Notebook
Data: kaggle
Presentation:

THE UNIVERISTY OF TEXAS AT AUSTIN AI in Healthcare

Explainable AI for Dermatology: Combining CNN Classification, Grad-CAM, and Vision–Language Models

Aaron Rodriguez

AI in Healthcare High-Risk Project, The University of Texas at Austin, aaron.n.rodriguez@utexas.edu

Skin cancer, and more so melanoma, is a serious public health issue because of its aggressive behavior and importance of early detection to guarantee a better patient outcome. The goal of this project was to ascertain if a hybrid deep learning pipeline of a vision classification model and vision–language model (VLM) can aid in the assessment of lesions and produce descriptive captions for medical images. A MobileNetV2 classifier was utilized to classify a dermoscopic image dataset into melanoma, benign keratosis, and melanocytic nevi. Grad-CAM visualizations were utilized for model explanation and prediction grounding. The visual explanations were fed to an InstructBLIP-based VLM for generating region-based captions, which were assessed against lesion-specific vocabularies. A review of more than 200 images concluded that although the classifier itself performed comparatively well, the VLM captions were frequently incomplete or misleading and their value as stand-alone clinical tools limited. Workflow, by contrast, illustrates that it is possible to combine interpretability and descriptive AI in dermatology and sets the stage for continued refinement by fine-tuning and increased-capacity models.

CCS CONCEPTS • Computing Methodologies

Additional Keywords and Phrases: Deep Learning, Computer Vision, Vision Language Models, dermoscopic

1 INTRODUCTION

Skin cancer is still among the most prevalent cancers in the world; early detection significantly enhances survival. In most of the globe, specialist dermatological examinations are not readily available and even experienced physicians are unable to distinguish between benign and malignant lesions when lesions are somewhat similar. In such types of environments, **computer vision** systems provide the potential to screen large numbers of dermoscopic images and flag suspicious lesions for specialist assessment. Even with the progress made in convolutional neural networks (CNNs), standard image classifiers are akin to black boxes and do not offer any explanation of the result. This lack of transparency hinders clinical adoption since doctors must understand why an AI algorithm arrived at a conclusion.

Recent work has explored **explainable AI (XAI)** techniques for medical images. As an example, researchers have utilized gradient-weighted class activation mapping **(Grad-CAM)** and similar saliency methods to identify areas of an image responsible for classifying skin lesions and shown that these marked areas of the image tend to overlap with dermatologist annotations. **Vision-language models (VLMs)** for medical images are supporting explainability. VLMs can describe an image in free text and can return captions that identify clinically significant features.

This paper examines if accurate predictions and interpretable outputs for dermoscopic images of skin lesions are achievable through the integration of a CNN classifier, Grad-CAM and a VLM. We establish a proof-of-concept system on the HAM10000 dermatoscopic dataset, train a CNN classifier to classify images as melanoma (mel), nevus (nv) or benign keratosis-like lesion (bkl), use Grad-CAM to highlight the salient areas, and ask InstructBLIP to produce descriptions of the lesion. We then evaluate the classifier, examine Grad-CAM maps, and investigate VLM captions both qualitatively and quantitatively. Although the final captions are generally generic and misclassify minority classes, the effort is a high-risk exploratory pursuit in explainable AI.

2 RELATED WORK

Explainable AI techniques have been widely utilized in skin lesion classification. Nunnari et al. [2021] proposed a ResNet-based network for skin tumor detection and used Grad-CAM to visualize saliency maps against dermatologist labels. Their outcome of saliency maps aligned with annotated areas, thereby simplifying interpretation. Shah et al. [2024] compared CNNs with Grad-CAM and occlusion sensitivity on the HAM10000 dataset and demonstrated that visual explanations helped clinicians understand model failure examples.

Compared to these papers, our study integrates a standard MobileNetV2 classifier with Grad-CAM and the InstructBLIP vision-language model to check whether free-text descriptions contain class discriminative information. Unlike other research, our focus is not state-of-the-art accuracy but rather exploring the potential and boundaries of applying a classifier, saliency maps, and an general purpose VLM in dermatology.

3 METHODOLOGY

3.1 Data

We utilized the **HAM10000** dataset from the International Skin Imaging Collaboration (ISIC). The database consists of 10,015 dermatoscopic images labeled as melanoma (mel), melanocytic nevus (nv) or benign keratosis-like lesion (bkl) with meta-data including age, sex and lesion site. Upon filtering into the three classes, the class distribution was highly imbalanced: 6,705 images (75.2 %) were nv, 1,113 (12.5 %) mel and 1,099 (12.3 %) bkl. A stratified train/validation split was utilized, and images resized to 224 × 224 pixels and normalized to ImageNet channel statistics. The training set was augmented with random crops, horizontal flips and small rotations.

3.2 Classifier Training

We chose MobileNetV2, a light CNN model pre-trained on ImageNet, and fine-tuned its classification head while freezing the convolutional backbone. The network was trained for 10 epochs with cross-entropy loss and Adam optimizer with a learning rate of 10^{-4} and batch size 32. Early stopping and learning-rate reduction on plateau were employed to avoid overfitting of the network. Figure 1 is the classification report and confusion matrix of the refined model, where overall validation accuracy is 74.7%. Although performance was better for mel class than in previous runs, class wise recall was still imbalanced: recall for mel was 81.2%, bkl was 76.8%, and nv was 73.2%. The confusion matrix illustrates that nv continues to significantly outperform correct predictions but misclassifications of mel decreased, which suggests that hyperparameter tuning did influence minority class identification. Targeted additional augmentation and class balancing can fill the remaining gap in performance between classes.

Classification Report:

	precision	recall	f1-score	support
bkl	0.6377	0.7682	0.6969	220
mel	0.3694	0.8117	0.5077	223
nv	0.9543	0.7323	0.8287	1341
accuracy			0.7466	1784
macro avg	0.6538	0.7707	0.6778	1784
weighted avg	0.8422	0.7466	0.7723	1784

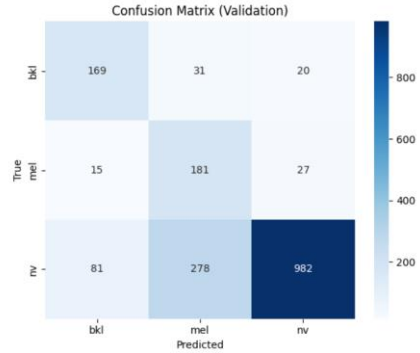

Figure 1: Classification report and confusion matrix on the 1,784 image validation set.

3.3 Explainability via Grad-CAM

To understand which image regions the CNN relied on, we implemented **Grad-CAM**. We calculated the gradient of class logit with respect to feature maps of last conv layer for a given image and target class, spatially averaged gradients to get weights, and then multiplied them with feature maps to produce a coarse heatmap. The heatmap was then normalized, resized to input size, and overlaid on the original image to mark significant regions. Examples are depicted in Figure 2: the model properly focuses on the body and boundary of the lesion and ignores the hair and background; examples display that heatmaps are consistent across diverse nevus cases.

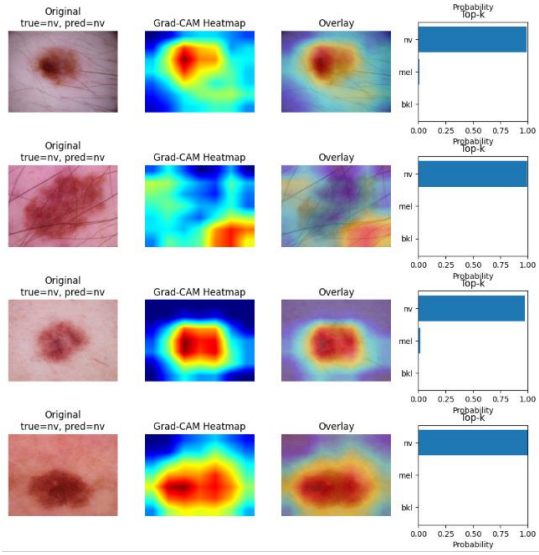

Figure 2: Grad-CAM heatmaps for different images (rows). Columns show the original image, Grad-CAM heatmap, overlay, and top-k class probabilities.

3.4 Vision Language Model Descriptions

We explored whether a general-purpose vision–language model can generate descriptive captions of lesions. We used **InstructBLIP**, an instruction-tuned version of BLIP that accepts both images and prompts. Three prompts were used:

- **GLOBAL:** "You are a expert in examining images of lesions. Provide one clinical sentence. State symmetry (symmetric/asymmetric), border (regular/irregular), color (single/multiple), texture (flat/raised, scaly/smooth). No diagnosis."
- **ROI:** "Focus only on the lesion. One clinical sentence stating symmetry, border, color distribution, diameter in millimeters if visible, and texture. No diagnosis."
- **FOCUSED:** "With background dimmed, describe symmetry, border, color distribution, texture, and any visible structures (dots, streaks, network). No diagnosis."

For each image, we first computed Grad-CAM and extracted the **region of interest (ROI)** by thresholding the heatmap and cropping a bounding box around the highest importance region. We then fed the original, ROI and background dimmed images to InstructBLIP with the corresponding prompts. Figure 3 shows the pipeline: the original image passes through the classifier, Grad-CAM produces a heatmap and ROI crop, and the VLM generates textual captions.

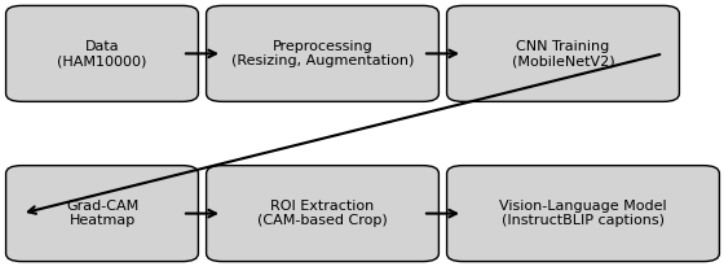

Figure 3: Workflow of the explainable AI pipeline: data passes through preprocessing and CNN training; Grad-CAM highlights salient regions; an ROI is cropped for the Vision-Language Model which produces clinical captions.

3.5 Caption-Derived Classification Experiment

To assess whether captions contain class discriminative properties, we implemented a **lexicon-based mapping**. We constructed keyword sets that characterized melanoma (*asymmetric*, *irregular border*, *multiple colors*), keratosis (*waxy*, *sharply demarcated*, *keratotic*) and nevi (*uniform*, *regular border*, *even color*). For each caption, we determined the number of occurrences of these keywords and used the most common class. We applied this mapping to the **ROI captions** for 200 randomly sampled images, guiding Grad-CAM by the true class.

4 RESULTS AND DISCUSSION

4.1 Classifier Performance

As evident in Figure 1, the classifier had 74.7% validation accuracy but class-skewed recall: 73.2% for nevi, 76.8% for keratosis, and 81.2% for melanoma. While these values indicate a significant improvement in melanoma detection compared to earlier runs, they also illustrate persistent class specific variation. This variability may stem from inherent class imbalance in the dataset (nevus comprises most images) and overlapping visual features between classes. Continued

efforts in class balancing, targeted augmentation, or specialized loss functions could help further balance recall across all categories.

4.2 Grad-CAM Insights

The Grad-CAM overlays (Figure 2) demonstrate that the model focuses on the lesion body and border. In correctly classified nevi, heatmaps concentrate on the symmetric central region and ignore hairs; in misclassified or ambiguous cases, attention sometimes spreads to background artefacts, explaining incorrect predictions. This interpretability enables clinicians to verify whether the model attends relevant regions and to identify cases requiring caution.

4.3 Vision-Language Model Outputs

Qualitative inspection of the VLM captions shows that they often capture basic visual features such as color and symmetry but rarely provide diagnostic detail. For example:

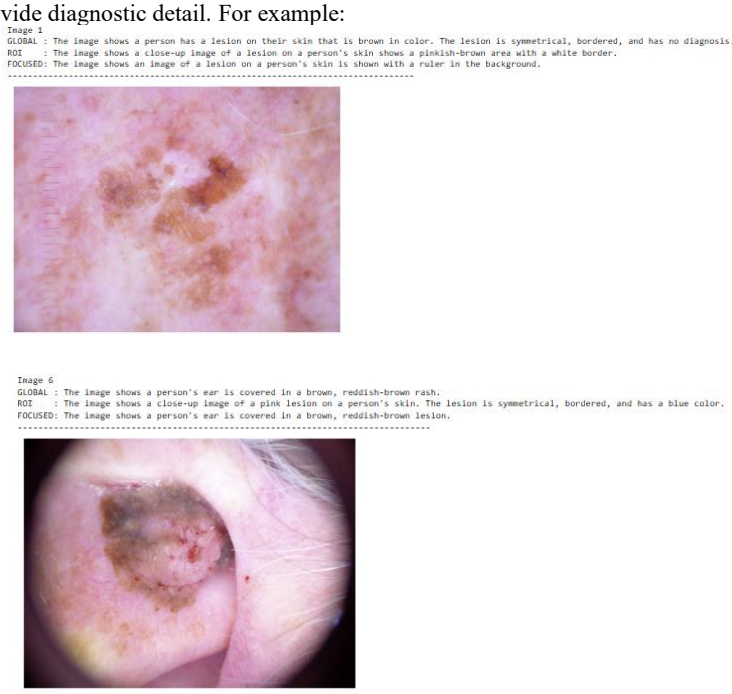

Figure 4: Vision language model output examples.

Figure 4 – Image 1: "The image shows a lesion on their skin that is brown in color. The lesion is symmetrical, bordered, and has no diagnosis." (Global)

Figure 4 – Image 6: "The image shows a person's ear covered in a brown, reddish-brown lesion." (Global)

These statements properly mention colors and symmetry but leave out irregular borders or variegated pigmentation. The ROI and focused captions occasionally mention the white border or diameter but add on irrelevant details (*ruler in the background*). This reflects the model's training on general image captions rather than dermatology reports.

4.4 Caption Derived Classification

Figure 5 summarizes the lexicon-based experiment on 200 images. The captions predicted the majority class **nv** for 80 % of images. Precision and recall for melanoma and keratosis were near zero; the macro-averaged F1 score was 0.260. This demonstrates that captions are **not reliably class discriminative** despite being grounded in the image.

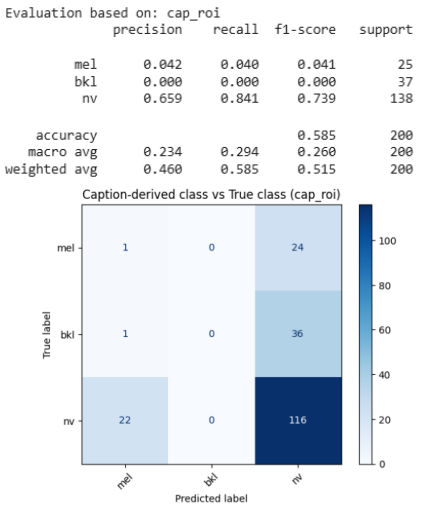

Figure 5: Classification report and confusion matrix for caption-derived class labels (ROI captions) vs true class on 200 images.

5 CONCLUSION AND FUTURE WORK

This research investigated a new combination of a CNN classifier, Grad CAM visualization and a Vision–Language Model for dermoscopic image analysis. Though the MobileNetV2 classifier showed acceptable overall accuracy, its recall for melanoma was poor, identifying a glaring weakness working with imbalanced datasets. Transparent overlays in the use of Grad CAM visualizations aided understanding of the focus of the network and emphasized difficult cases with bleeding of attention into non-salient areas. The InstructBLIP explanations provided simple descriptive stories, but a lexicon evaluation found that such explanations contained non-discriminative clinical details and skewed towards the majority class.

Despite modest results, the project demonstrates high effort and risk-taking by combining multiple AI components. Future work could improve performance and interpretability by:

1. **Class re-balancing and loss adjustments** (e.g., focal loss, oversampling) to address the low melanoma recall.
2. **Domain-specific fine-tuning** of the VLM using dermatology report corpora and curated lesion descriptions, encouraging the model to mention diagnostic attributes such as asymmetry, irregular borders and variegated pigmentation.
3. **Advanced saliency methods**, such as Grad-CAM++ or Score-CAM, to produce more precise heatmaps and reduce background artefacts.
4. **Structured caption evaluation** against radiologist-written descriptions to quantify alignment between AI-generated narratives and expert observations.

Ultimately, this high-risk project illustrates the difficulties of combining classification, explainability, and natural language generation in medical imaging. This paper lays the groundwork for future improvements towards clinically useful multi-modal AI systems.

REFERENCES

[1] Nunnari, F., Kadir, M.A., Sonntag, D. (2021). On the Overlap Between Grad-CAM Saliency Maps and Explainable Visual Features in Skin Cancer Images. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2021. Lecture Notes in Computer Science(), vol 12844. Springer, Cham. https://doi.org/10.1007/978-3-030-84060-0_16

[2] Shah, S. A. H., Shah, S. T. H., Khaled, R., Buccoliero, A., Shah, S. B. H., Di Terlizzi, A., Di Benedetto, G., & Deriu, M. A. (2024). Explainable AI-Based Skin Cancer Detection Using CNN, Particle Swarm Optimization and Machine Learning. *Journal of imaging*, *10*(12), 332. https://doi.org/10.3390/jimaging10120332

[3] Dai, W., Li, J., Li, D., Tiong, A.M., Zhao, J., Wang, W., Li, B.A., Fung, P., & Hoi, S.C. (2023). InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *ArXiv, abs/2305.06500*.