
Toward Trustworthy AI for Medical Imaging: ResNet-Based Diagnosis with Grad-CAM++ Explanations

Aryan Jagani

University of Maryland, Baltimore County (UMBC)
{ajagani1}@umbc.edu

Vishvakumar Patel

University of Maryland, Baltimore County (UMBC)
{vpate119}@umbc.edu

Abstract

Deep learning models can perform well in medical image classification, but their adoption in clinical settings is often limited because they lack clear and trustworthy explanations. In critical situations like skin lesion diagnosis, clinicians need accurate predictions and clear justifications for the model's decisions. We present a ResNet-50 based skin lesion classifier enhanced with explainable AI methods, focusing on Grad-CAM++ to provide detailed, class-specific heatmaps. Using the ISIC 2019 dermoscopic dataset, we compare Grad-CAM, Grad-CAM++, and the upcoming Score-CAM across different network depths. We also add a simple text-based explanation tool that summarizes important areas in terms that are easy to understand. Our findings show that Grad-CAM++ gives clearer and more lesion-focused explanations at deeper layers. This supports our goal of creating more reliable medical imaging systems.

1 Introduction

Skin cancer is among the most frequently diagnosed types of cancer and a serious public health problem. The survival chances in skin cancer, especially in melanoma, increase greatly if it is diagnosed at an early stage. Dermoscopy or dermoscopic imaging is a technique in dermatology which has become a fundamental part of dermatology. The drawback in dermoscopic image analysis for accurate skin cancer diagnosis is the need for expertise, which can have variations among observers. Hence, computerized systems for diagnosis are a reliable option.

Recently, breakthroughs in deep learning, including convolutional neural networks (CNNs), have achieved major breakthroughs in medical image analysis. ResNet architectures have shown great efficacy in skin lesion image classification and have even approached human-level dermatologist performance in certain studies. Such algorithms can obviously learn robust associations with a variety of lesion characteristics, including color, texture, and morphology, from massive datasets such as ISIC without requiring any additional information. Although such algorithms have very high accuracy in making predictions, they are regarded as a black box since they do not give a clear explanation of lesion characteristics in a particular image used for making a prediction.

A major drawback in this case is a lack of interpretability. In medical settings, where trust and transparency are essential, predictive accuracy with flawed or unclear results can have serious repercussions for patients. A model, in this case, needs to be understood by a medical practitioner in such a way that it explains why a given predictive output was arrived at and whether this output is based on

meaningful information or mere correlations. Therefore, interpretability is an important consideration in using artificial intelligence in medical imaging.

To overcome this challenge, methods called Explainable Artificial Intelligence, or XAI methods, have been developed to allow humans to understand an image classifier's internal workings and decisions using deep learning models. To address image classification, methods have concentrated on localizing image regions with strong influence on model output. Among these methods are those named as Class Activation Map, which are widely used since they are able to create a heatmap image with strong capabilities of being interpreted by an expert. Heatmaps are most important in the field of dermatology because they allow a medical professional to inspect a marked area side by side with a lesion.

Among the various methods based on CAM, Grad-CAM is widely used because of its ease of implementation and usability with all architectures of CNN. Although, in comparison with other methods, Grad-CAM explanations can be a bit more coarse-grained and, in some instances, include regions in the image where there is no relevance clinically. To counter these issues, more advanced versions, such as Grad-CAM++ and Score-CAM, have been explored. Grad-CAM++ gives improved instance-localization capabilities and finer details, and Score-CAM is not bound by gradients and can provide more accurate explanations. A comparative analysis of these in a medical imaging setting is required in order to have a complete insight into their capabilities and weaknesses.

In this work, a skin lesion image classifier using a ResNet50 model based on images provided in the ISIC 2019 dataset, which comprises a variety of benign and malign skin lesions, is proposed. To assess the different skin lesion regions emphasized by different explanation techniques, a variety of XAI algorithms such as Grad-CAM, Grad-CAM++, and Score-CAM, which employ a variant of the class activation mapping approach called CAM, will be used. Moreover, a brief human-understandable text description summarizing the model's attention regions and rationale for predictions will be produced. In summary, this research can be considered relevant in the context of improving deep learning methods for classifying skin lesions and making these systems more transparent. Additionally, this research can help improve the trustworthiness of AI systems and make them more suitable for integration in a medical setting.

2 Problem Statement and Hypothesis

2.1 Problem Statement

Among deep learning algorithms, convolutional neural network models have proved to be very accurate in classifying skin lesions using dermoscopy images. ResNet-50 models have the capability to learn complicated visual patterns in images and can classify them with an accuracy comparable to a dermatologist. Meanwhile, these models have remained black boxes despite their accuracy in making predictions.

A major obstacle in using AI-powered diagnostic systems in a medical setup is a lack of interpretability. Noisy or non-interpretable results can sometimes be dangerous because medical practitioners need to check if a particular region in a lesion, which is relevant to a medical conclusion, is being used for taking a particular decision, rather than focusing on some artifacts in the background. A major problem with existing solutions, such as Grad-CAM, is that they provide a visual explanation in the form of a heatmap, which might not be very precise.

Hence, a need arises for a thorough assessment of different state-of-the-art methods of class activation mapping, such as Grad-CAM, Grad-CAM++, and Score-CAM, in the realm of skin lesion segmentation and classification. Moreover, relying solely on visualization methods for better understanding and interpretation may not be sufficient, thus pointing towards a need for additional methods of providing human-readable explanations in this realm.

2.2 Hypothesis

2.2.1 Primary Hypothesis

Advanced Class Activation Mapping (CAM)-based explainability methods, such as Grad-CAM++ and Score-CAM, will provide more precise and medically relevant localization of skin lesion regions compared to standard Grad-CAM when applied to a ResNet-50 skin lesion classification model.

2.2.2 Secondary Hypothesis

The inclusion of concise, human-readable textual explanations alongside visual heatmaps will improve the interpretability and perceived trustworthiness of model predictions for clinical decision support.

2.2.3 Performance Consistency Hypothesis

Incorporating explainability techniques will not significantly degrade the classification performance of the ResNet-50 model on the ISIC 2019 dataset, allowing interpretability to be enhanced without sacrificing predictive accuracy.

3 Related Work

Grad-CAM, introduced by Selvaraju et al. (2), enabled class-discriminative localization in convolutional neural networks by weighting feature maps using gradients. While effective, Grad-CAM often produces diffuse heatmaps with limited spatial precision. Chattopadhyay et al. (3) proposed Grad-CAM++, which incorporates higher-order gradient information to improve localization, particularly in cases involving multiple salient regions.

Several studies have applied Grad-CAM-based methods to medical imaging. Yang et al. (4) demonstrated improved pneumonia diagnosis using ResNet combined with Grad-CAM, but relied primarily on qualitative visualization. Recent work emphasizes that visual plausibility alone is insufficient; explanation faithfulness must be quantitatively evaluated using metrics such as Intersection over Union (IoU), pointing-game accuracy, and deletion or insertion curves.

Beyond gradient-based methods, Score-CAM offers a perturbation-based, gradient-free alternative that produces smoother saliency maps at higher computational cost (5). Additionally, the reliability of predictions themselves is critical in clinical AI, motivating the use of calibration techniques such as temperature scaling and Expected Calibration Error (ECE). Robustness analyses further highlight that explanations should remain stable under model or input perturbations.

Our work builds on this literature by systematically comparing Grad-CAM and Grad-CAM++ across network depth, outlining quantitative evaluation strategies, and augmenting visual explanations with clinician-friendly textual summaries.

4 Methodology

4.1 Dataset: ISIC 2019

We evaluate our approach using the ISIC 2019 skin lesion classification dataset (6), a large-scale public benchmark consisting of dermoscopic images annotated into nine diagnostic categories: MEL (melanoma), NV (nevus), BCC (basal cell carcinoma), AK (actinic keratosis), BKL (benign keratosis-like lesions), DF (dermatofibroma), VASC (vascular lesions), SCC (squamous cell carcinoma), and UNK (unknown).

In our pipeline, the dataset is downloaded programmatically using `kagglehub`, ensuring reproducibility and eliminating manual preprocessing steps. Ground-truth labels are provided in one-hot encoded format across the nine diagnostic classes. We load both the label file and associated metadata and merge them using the shared image identifier (`image`) via an inner join to guarantee label-image consistency. To convert the one-hot encoding into a single class label suitable for supervised learning, we select the class with the maximum value using `idxmax`. Class names are then mapped to integer labels using a fixed `class_to_label` dictionary, which remains consistent throughout training and evaluation.

	image	age_approx	anatom_site_general	lesion_id	sex	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	UNK	class_name	label
0	ISIC_0000000	55.0	anterior torso		NaN	female	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	NV	1
1	ISIC_0000001	30.0	anterior torso		NaN	female	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	NV	1
2	ISIC_0000002	60.0	upper extremity		NaN	female	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	MEL	0
3	ISIC_0000003	30.0	upper extremity		NaN	male	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	NV	1
4	ISIC_0000004	80.0	posterior torso		NaN	male	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	MEL	0

Figure 1: Merged ISIC 2019 dataset sample with metadata and labels.

The table below shows a representation of the resulting combined dataset after merging both ISIC 2019 metadata and ground truth label files based on a common image identifier. The resulting combined dataframe consists of a total of 25,331 samples with 14 attributes.

A row in this table corresponds to a lesion image. The image column holds a distinct ISIC image id. Other demographic and clinical information available are age, sex, and general anatomical site. The lesion id in this table can have missing values in case a lesion grouping is not recorded for an image.

The diagnostic labels are presented in a one-hot encoded form for nine classes: MEL (melanoma), NV (naevus), BCC (basal cell carcinoma), AK (actinic keratosis), BKL (benign keratosis-like lesions), DF (dermatofibroma), VASC (vascular lesions), SCC (squamous cell carcinoma), and UNK (unknown). To conduct a supervised learning task, a one-hot vector is reduced to a single class. The class name is given in the class name column, and in the label column, an integer class representation is given.

In summary, this combined and processed dataframe serves as a common input source for both dividing datasets, model learning, and analysis of model explanations.

4.2 Preprocessing and Data Augmentation

All images are processed to satisfy the input requirements of ResNet-50. Images are loaded using the Python Imaging Library (PIL), converted to RGB format, and transformed into PyTorch tensors. We standardize the spatial resolution to 224×224 pixels and normalize pixel intensities using ImageNet mean and standard deviation values (mean [0.485, 0.456, 0.406], standard deviation [0.229, 0.224, 0.225]), matching the statistics used during ImageNet pretraining.

To improve generalization and mitigate overfitting, we apply data augmentation exclusively during training. Augmentations include `RandomResizedCrop(224)` to introduce scale variability, random horizontal and vertical flips to enhance orientation invariance, mild `ColorJitter` to account for illumination and color variation in dermoscopic imaging, and `RandomRotation(15)`° to model small rotational perturbations. Validation preprocessing is deterministic, using `Resize(256)` followed by `CenterCrop(224)`, ensuring that evaluation results are stable and comparable across runs.

4.3 Train-Validation Split

We partition the dataset into training and validation subsets using an 80/20 split. To preserve the highly imbalanced class distribution characteristic of dermoscopic datasets, we perform stratified sampling based on the integer class labels. This ensures that minority classes are represented proportionally in both subsets and prevents biased performance estimates.

4.4 Model Architecture and Training

We adopt a ResNet-50 convolutional neural network pretrained on ImageNet as the backbone for skin lesion classification. The final fully connected layer is replaced with a new linear classification head producing logits for the nine ISIC classes. All layers are fine-tuned end-to-end.

Training is performed using cross-entropy loss and the Adam optimizer with a learning rate of 1×10^{-4} . We use a batch size of 64 and train for 10 epochs. To improve computational efficiency and reduce GPU memory consumption, we enable automatic mixed precision (AMP) training using `torch.cuda.amp.autocast` in conjunction with `GradScaler`. Each training iteration consists of

a forward pass under autocast, loss computation, scaled backpropagation, and an optimizer update with dynamic gradient scaling. Upon completion of training, the model weights are saved to disk as `isic_resnet50.pth`.

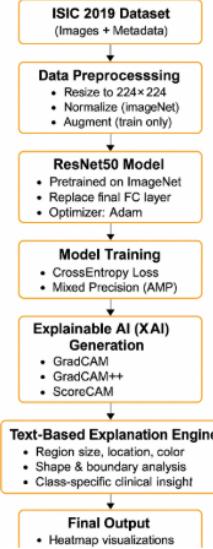


Figure 2: Overall workflow of the proposed skin lesion classification and explainability framework.

4.5 CAM-Based Explainability Methods

To interpret model predictions, we generate class-discriminative saliency maps using CAM-based explainability techniques implemented via the `pytorch-grad-cam` library. For each input image, we compute softmax probabilities and select the predicted class. A `ClassifierOutputTarget` corresponding to this class is used to ensure that the explanation reflects the model’s actual decision rather than a user-specified label.

We compare three explainability methods:

- **Grad-CAM**, which computes importance weights by globally averaging the gradients of the target class with respect to convolutional feature maps;
- **Grad-CAM++**, which incorporates higher-order gradient information to better handle multiple or overlapping salient regions;
- **Score-CAM**, a gradient-free method that estimates importance scores through activation-guided perturbations, producing smoother but more computationally expensive saliency maps.

4.6 Layer-Wise Saliency Analysis

To analyze how interpretability evolves with network depth, we extract CAMs from multiple stages of the ResNet-50 architecture. Specifically, we generate explanations from the final blocks of `Layer1`, `Layer3`, and `Layer4`. Shallow layers capture low-level edges and textures, while deeper layers encode increasingly abstract and class-discriminative representations. This layer-wise comparison enables systematic analysis of how lesion localization sharpens as features become more semantic.

4.7 Heatmap Post-processing and Visualization

Raw CAM outputs are min–max normalized and optionally smoothed using a Gaussian blur to reduce noise. Heatmaps are then resized to the input image resolution and overlaid onto the original RGB image using `show_cam_on_image`. These overlays form the basis of all qualitative comparisons presented in the results section.

4.8 Text-Based Explanation Engine

In addition to visual explanations, we introduce a rule-based text explanation engine to summarize salient regions in human-readable terms. Using the Grad-CAM++ heatmap, we first threshold the normalized saliency map at $0.6 \times \max(\text{CAM})$ to isolate highly activated regions. Connected component analysis is performed, and the largest salient region is selected as the primary focus area.

From this region, we compute interpretable geometric descriptors, including spatial location (upper/lower and left/right based on centroid position), relative size (small, medium, or large based on area fraction), and shape (circular versus elongated based on eccentricity). These descriptors are assembled into a structured natural-language template. Finally, a brief class-specific clinical note is appended using a fixed rule dictionary, yielding a concise explanation that accompanies the predicted class label and confidence score.

5 Results and Discussion

5.1 Classification Performance

The trained model achieves 81% training accuracy and 82% validation accuracy. Performance is strongest on frequent classes (e.g., NV and BKL) and weaker on rare classes (e.g., MEL and SCC), consistent with the observed class imbalance.

5.2 Qualitative XAI Comparison

We compare Grad-CAM and Grad-CAM++ across network depth:

- **Shallow layers (Layer1):** heatmaps highlight edges and textures and are not lesion-specific.
- **Middle layers (Layer3):** heatmaps begin to align with lesion boundaries and structure.
- **Deep layers (Layer4):** heatmaps become most clinically relevant, localizing lesion regions used for classification.

Across layers, Grad-CAM++ produces sharper, more focused heatmaps than Grad-CAM, while Score-CAM (planned) is expected to yield smoother maps at higher computational cost.

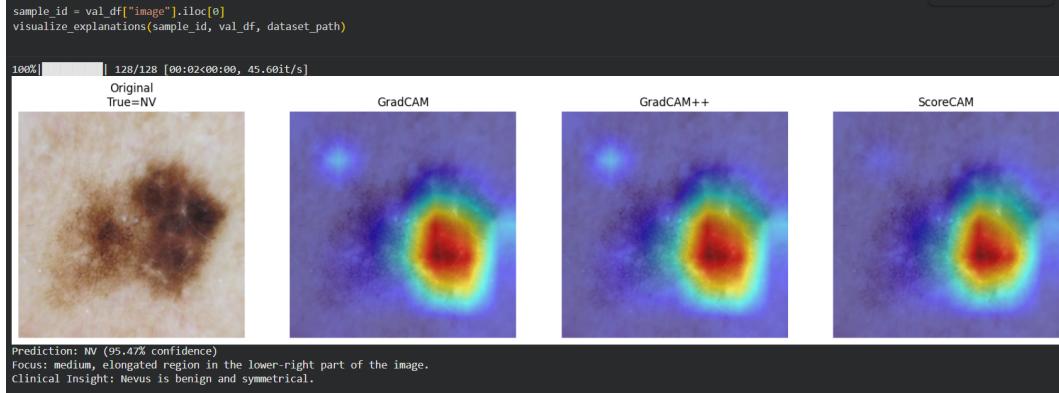


Figure 3: Qualitative comparison of explainability methods for a benign nevus (NV) case. The original dermoscopic image (left) is followed by Grad-CAM, Grad-CAM++, and Score-CAM heatmaps generated from the same ResNet-50 prediction. Grad-CAM produces a broader, less localized explanation, while Grad-CAM++ yields a sharper and more lesion-focused activation aligned with clinically relevant regions. Score-CAM generates smoother saliency maps at increased computational cost. The model predicts NV with 95.47% confidence, and the accompanying text-based explanation summarizes the salient region as a medium-sized, elongated area in the lower-right portion of the lesion, consistent with benign nevus characteristics.

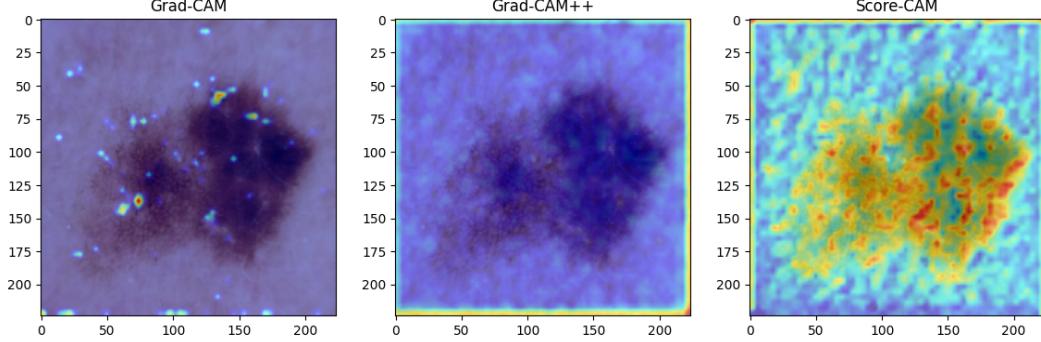


Figure 4: Early convolutional layer (Layer1) explanations highlight low-level textures and scattered activations. This figure compares three CAM-based explanation methods—Grad-CAM, Grad-CAM++, and Score-CAM—generated from an early convolutional layer (Layer 1) of the ResNet-50 model. At this stage, the network primarily captures low-level visual features such as edges, texture variations, and color contrasts rather than lesion-level semantics. Consequently, the explanations appear scattered and texture-focused, rather than clearly outlining the lesion.

Grad-CAM: Grad-CAM produces sparse and fragmented activations distributed across the image. The highlighted regions often correspond to high-contrast texture areas rather than coherent lesion structures.

Grad-CAM++: Grad-CAM++ generates smoother and more spatially coherent heatmaps compared to Grad-CAM. However, the activations remain focused on general texture patterns and may include edge-related artifacts.

Score-CAM: Score-CAM produces the most diffuse heatmap, covering a broader portion of the lesion region. This reflects the influence of multiple low-level activation maps at this early stage.

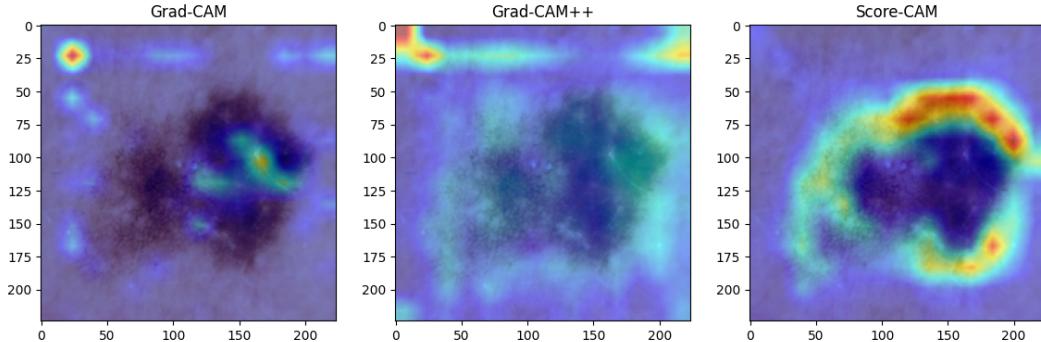


Figure 5: Intermediate layer (Layer3) explanations begin to capture lesion structure and spatial organization. This figure illustrates CAM-based explanations generated from an early convolutional layer (Layer 1) of the ResNet-50 model using Grad-CAM, Grad-CAM++, and Score-CAM. At this stage, the network primarily captures low-level visual features such as edges, texture variations, and color contrasts rather than lesion-level semantics. Grad-CAM produces sparse and fragmented activations, Grad-CAM++ generates smoother but still texture-focused heatmaps, and Score-CAM yields broader activations influenced by multiple low-level feature maps. Consequently, the explanations appear scattered and texture-oriented, without clearly outlining the lesion structure.

Grad-CAM: At the intermediate layer, Grad-CAM begins to highlight more structured regions within the lesion. The activations are less scattered and start aligning with meaningful lesion features, though some background sensitivity remains.

Grad-CAM++: Grad-CAM++ produces smoother and more spatially organized heatmaps, capturing broader lesion structures while reducing fragmented activations observed in earlier layers.

Score-CAM: Score-CAM shows strong, well-defined activations around prominent lesion regions, indicating improved spatial localization and a clearer focus on lesion structure at this intermediate stage.

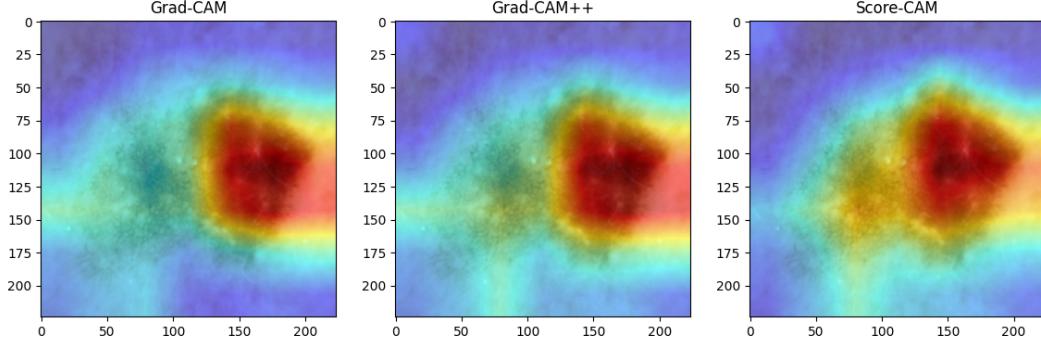


Figure 6: Deep layer (Layer4) explanations provide the most clinically meaningful localization, focusing on lesion regions critical for classification. Across all layers, Grad-CAM++ produces sharper and more spatially consistent heatmaps compared to Grad-CAM, while Score-CAM generates smoother but less localized activations.

Grad-CAM: Grad-CAM highlights the primary lesion region with strong activations, indicating that the model relies on high-level lesion features for classification. However, the heatmap boundaries remain relatively broad.

Grad-CAM++: Grad-CAM++ produces sharper and more spatially consistent activations concentrated on clinically relevant lesion regions, offering improved localization compared to Grad-CAM.

Score-CAM: Score-CAM generates smoother activations that broadly cover the lesion area. While the focus remains on relevant regions, the localization is less precise than Grad-CAM++.

5.3 Interpretability via Text Summaries

The text-based explanation engine provides concise descriptions of region size, location, shape, and boundary properties, making it easier to interpret heatmaps without solely relying on visual inspection.

6 Limitations

- **Class imbalance:** minority malignant classes remain difficult; improved balancing is needed.
- **Quantitative faithfulness:** while we outline IoU, pointing-game, and deletion curves, broader quantitative evaluation remains incomplete.
- **Post-hoc explanations:** CAM methods do not guarantee causal faithfulness and can still highlight correlated artifacts.
- **Metadata underuse:** ISIC metadata (age, sex, anatomical site) was not fully leveraged in this pipeline.

7 Future Work

We are going to: (i) compare different architectures (EfficientNet, Vision Transformers) to see which one works best, (ii) enhance the performance of the minority class by reweighting/oversampling and

using a better augmentation, (iii) completely integrate Score-CAM and locally implement quantitative XAI metrics (IoU, pointing-game, deletion/insertion curves), (iv) clearly show calibration analysis by adding it, and (v) develop an interactive interface (e.g., Streamlit) for real-time visualization and getting clinician feedback.

8 Conclusion

We presented a ResNet-50 skin lesion classifier with CAM-based explainability. Layer-wise analysis indicates that deeper layers produce the most clinically relevant localization. Among CAM methods, Grad-CAM++ yields sharper, more focused heatmaps than Grad-CAM, and a text-based explanation engine further improves interpretability. Together, these components contribute toward more trustworthy AI for medical imaging.

Team Contributions (Post-Conclusion)

Aryan Jagani: model architecture and training pipeline, Grad-CAM++ implementation, calibration analysis (planned/partially explored), literature review and method comparison design, and overall integration of XAI outputs.

Vishvakumar Patel: dataset preprocessing and augmentation, segmentation alignment (where applicable), explainability metric planning/implementation scaffolding, robustness analysis contributions, and visualization support.

Shared: documentation, experiment organization, analysis synthesis, and presentation preparation.

References

- [1] Samek, Wojciech Wiegand, Thomas Müller, Klaus-Robert. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*. 1-10. 10.48550/arXiv.1708.08296.
- [2] Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *International Journal of Computer Vision*, vol. 128, no. 2, Oct. 2019, pp. 336–359. Springer Science and Business Media, doi:10.1007/s11263-019-01228-7.
- [3] Chattopadhyay, Aditya Sarkar, Anirban Howlader, Prantik Balasubramanian, Vineeth. (2017). Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. 10.48550/arXiv.1710.11063.
- [4] Yang, Yuting Mei, Gang Piccialli, Francesco. (2021). Explainable Deep Learning Models on the Diagnosis of Pneumonia. 10.1109/CHASE52844.2021.00032.
- [5] H. Wang et al., "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 111-119, doi: 10.1109/CVPRW50498.2020.00020.
- [6] Gutman, David; Codella, Noel C. F.; Celebi, Emre; Helba, Brian; Marchetti, Michael; Mishra, Nabin; Halpern, Allan. "Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)". eprint arXiv:1605.01397. 2016.

A Appendix: Suggested Quantitative XAI Metrics (Planned)

This appendix describes metrics intended for more rigorous explanation validation:

- **Intersection over Union (IoU):** overlap between thresholded heatmaps and ground-truth lesion masks (if segmentation is available).
- **Pointing-game accuracy:** whether the maximum saliency point falls inside the lesion region.
- **Deletion curves:** progressively remove most-salient pixels and measure confidence drop.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract/introduction summarize the ResNet-50 classification pipeline, Grad-CAM++ emphasis, layer-wise analysis, and the text-based explanation module.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Limitations are discussed in the *Limitations* section, including class imbalance and post-hoc faithfulness.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This work is empirical and does not present formal theoretical results.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results?

Answer: [Yes]

Justification: Dataset, preprocessing, split strategy, optimizer, learning rate, batch size, and environment are provided in Methods and Experimental Setup.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to reproduce results?

Answer: [No]

Justification: The dataset is public (ISIC 2019), but code release was not included as part of the course deliverables.

6. Experimental setting/details

Question: Does the paper specify all the training and test details necessary to understand the results?

Answer: [Yes]

Justification: The Experimental Setup section lists split, image size, optimizer, learning rate, loss, batch size, and compute environment.

7. Experiment statistical significance

Question: Does the paper report error bars or other appropriate information about statistical significance?

Answer: [No]

Justification: We report accuracy on a single split; repeated runs and confidence intervals are left to future work.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on compute resources?

Answer:

Justification: We specify Google Colab GPU usage; detailed runtime and GPU model are not reported.

9. Code of ethics

Question: Does the research conform with the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: We use a public dataset and do not deploy clinical decisions; we discuss limitations and risks.

10. Broader impacts

Question: Does the paper discuss positive and negative societal impacts?

Answer: [Yes]

Justification: The paper highlights clinical benefits and risks (miscalibration, misleading explanations) in Limitations/Future Work.

11. **Safeguards**

Question: Does the paper describe safeguards for responsible release of high-misuse-risk data/models?

Answer: [NA]

Justification: We do not release a model as a product; this is a course project prototype.

12. **Licenses for existing assets**

Question: Are original owners of assets credited and licenses respected?

Answer:

Justification: We credit ISIC 2019 and core methods; explicit license text is not included and should be added if required.

13. **New assets**

Question: Are new assets introduced well documented and documented alongside the assets?

Answer: [NA]

Justification: We do not introduce a new dataset; our text explanation rules are described in the Methods section.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing or human-subject research, are instructions/compensation provided?

Answer: [NA]

Justification: No crowdsourcing or human subject study was conducted.

15. **IRB approvals**

Question: Does the paper describe risks and IRB approvals for human subjects?

Answer: [NA]

Justification: No human subject study was conducted.

16. **Declaration of LLM usage**

Question: Does the paper describe usage of LLMs if important to core methods?

Answer: [NA]

Justification: LLMs are not part of the modeling or XAI pipeline.