Personal Section

"You can have 1,000 problems until you have a health problem. Then, you only have one problem." This saying used to define my life experience with eczema. It also represents the inspiration for my work with melanoma detection research.

Growing up in Shanghai, eczema found ways to affect my life in all aspects. I remember dreading summertime, when sweating was inevitable along with the unbearable itch that came with it. The condition also affected my relationship with physical activity. I was relatively athletic growing up, and having to deal with the burning on my skin after every sports game or practice sometimes felt like punishment for doing what I loved. Eating the wrong things or being stressed out could also lead to a breakout of rashes. Eczema was a part of my life that demanded attention in waves and fell into the background when it temporarily went away.

My relationship with eczema really changed when I moved to the U.S. for high school. I still don't know if it was a change in the climate or chemicals in the food or some other factor, but my condition worsened dramatically during a flareup that lasted a few months. For weeks I spent 2 hours scratching in bed every night because there was no way I could sleep. Showering and washing my hands became painful rituals. Having to explain my scratching to others constantly was also a struggle. During this time, it truly felt like my life was being controlled by my condition.

As I experimented with different treatments, ointments, and lifestyle changes, my symptoms began slowly fading again. I went through tons of different creams, prescriptions, and even different shampoos and laundry detergents to eliminate all possible sources of irritation. Eventually though, the prolonged flareup withdrew just as quickly as it had started.

It felt incredibly liberating for eczema to no longer be the one thing on my mind 24/7. It left an opening in my life that I quickly filled up with research during my junior year of high school. I started delving into papers and journals, reading about the causes of the condition and the most modern treatments. As a byproduct of reading about eczema, I learned a lot more about the field of dermatology in general. That's how I started learning more about melanoma, which caught my attention because of its deadly impact. While eczema is where I had started, I didn't see a clear path forward in the literature in terms of a gap I could target. With melanoma though, the frontier was being expanded with the most recent AI and technology developments.

I had been an avid programmer and engineer for years by that time, and it seemed like the perfect research project to combine my interest in dermatology and experience with coding. Over the course of over a year, I worked independently on my laptop with Google Colab computing resources to conduct experiments on melanoma detection AI models. When I wrote my paper, I checked in periodically with a science teacher from school that mentored me throughout the process.

Although I had previously worked with machine learning, it felt like I was starting with zero experience. I was coding in python, which is a language I am not at all familiar with. I had also never used Google Colab or any cloud computing environment before. I also had to familiarize myself with the different model architectures I was working with, recent developments in the literature, useful databases, and much more. Most importantly, I learned the value of understanding over implementation. My ability to reason and understand what the errors in my code meant was much more important than my knowledge of the syntax. Overall,

conducting my research project was a time of tremendous growth and learning for me. I had to solve my own technical issues, block out my time efficiently, and set goals for what I wanted to accomplish.

For students that are thinking about pursuing an independent research project: My biggest piece of advice is one I got from my mentor from school, and what I attribute the conception and success of my own work to: just pick something you're interested in and get started. Before I began, I didn't know what exact hypothesis I wanted to research or what test I wanted to conduct. I wasn't even sure if I wanted to commit to dermatology because there were other fields I was interested in. The truth is, commitment isn't a decision you can realistically make in a split second. Commitment is borne out of hours of reading and becoming an expert in your field, starting with a seed of inspiration and genuine curiosity. Don't be afraid to start researching about that first curiosity! In the worst case, you'll be a better student and a learner because of the skills you picked up and be able to pivot to another topic. In the best case, you'll continue your research and build an amazing project out of it.

Another important note is not to be scared to collaborate with others. This is something I struggled with during my project. As the old saying goes, alone you can go faster but together you can further. I spent many hours just poring over my code alone. The more I talked to my teachers, friends, and especially experts in the field though, I found new directions to take up in my research. When in doubt, reach out! The insights you pick up from others will be invaluable.

Research Section

The title of my research project is **"GradCAM Analysis Reveals Deep Learning Models Misclassify Benign-Appearing Melanomas Based on Possible Architectural Shortcomings Instead of Misdirection Attention"**. Here is a brief summary of my project in simpler terms:

I studied AI image recognition models that are trained to detect if a picture of a skin lesion is melanoma. I wanted to see how the model I trained would perform on melanoma samples that are very symmetrical and color uniform, which are usually characteristics of benign lesions. After testing the model, I used a tool called GradCAM to compare the model's performance on melanomas that were symmetrical and color uniform versus typical. GradCAM reveals the parts of an image that AI models focus on when making a classification decision.

Abstract

Despite accounting for only 1% of skin cancers, melanoma causes the vast majority of skin cancer deaths, with early detection increasing the five-year survival rate from 30% to over 99% (SEER, 2025). With the rapid development of artificial intelligence and image processing technology, deep learning algorithms have been applied extensively in healthcare to aid in the diagnosis and screening process for skin cancers (Li, 2022). While current models have achieved consistent performance on par with or surpassing clinicians (Pham, 2021), further classification research is hindered by the extreme morphological diversity of melanoma. This study evaluates the effect that uniformity and symmetry in melanoma samples have on model predictions and validates model attention using GradCAM analysis. Using the HAM10000 dataset and expert-drawn segmentation masks, color-uniformity and symmetry of the lesions were computed, allowing lesions exhibiting the top quartile of uniformity and symmetry to be identified as "benign-appearing" cases. An EfficientNet-B3 model trained on binary classification of melanoma was evaluated based on its performance for those challenging cases exhibiting benign features. As hypothesized, the evaluation revealed an 87.5% false negative rate for challenging cases versus 59.7% for typical presentations ($p = 0.065$). Surprisingly, GradCAM analysis revealed that misclassified challenging cases showed increased attention to lesion interiors (mean inside-fraction: 48.9% vs 24.6%, Cohen's $d = 1.107$), contradicting the hypothesis that models fail due to focusing on peripherals. These findings suggest that benign-appearing melanomas represent a critical blind spot in current AI systems, requiring targeted algorithmic improvements to prevent potentially fatal diagnostic delays.

I. Introduction

1. Deadly Impact of Melanoma

Melanoma of the skin, despite accounting for only 1% of skin cancers, causes the vast majority of skin cancer deaths and is one of the most common cancers in young adults (American Cancer Society 2025). Estimates show that more than 100,000 people will be diagnosed with melanoma in 2025, resulting in around 8,000 deaths (SEER 2025). Research shows that survival rates increase dramatically if the disease is detected early, with metastasized melanoma having a 34.6% 5-year relative survival rate, while localized cases have a 100% survival rate (SEER 2025). These statistics underscore the importance of early diagnosis followed up with proper treatment when approaching melanoma.

2. Difficulties in Melanoma Detection

Despite advances in clinical practices and technological assistance, melanoma of the skin remains one of the most challenging malignancies to detect. Studies show that even experienced dermatologists fail to detect 23.1% of melanomas (Chen, 2024). This is due to melanoma's vast morphological diversity, with many lesions appearing similar to common benign growths such as melanocytic nevi. A common systematic approach to diagnosis used by clinicians before confirming with biopsy is the ABCDE checklist, which assesses the lesions asymmetry, border irregularity, color, diameter, and evolution. While this method is effective for most typical melanoma presentations, malignant lesions can sometimes mimic benign conditions and easily evade detection.

3. Integration of AI models into Dermatology

Recent developments in AI have led to widespread integration of deep learning technology in the medical sector. Landmark studies such as Esteva et al. (2017) have proven that deep convolutional neural networks are especially effective when applied to image-based classification of skin cancers. Numerous other studies (Haddadin et al., 2024; Pham et al., 2021; Tschandl et al., 2021) have shown that diagnosis results improve with AI-clinician collaboration, or confirmed that AI can perform on par with or even better than expert dermatologists. However, aggregate performance metrics often obscure weaknesses with particularly challenging cases.

4. Benign-Appearing Melanomas

Benign-appearing melanomas, which exhibit high symmetry, color uniformity, and regular borders, represent a critical challenge for both human experts and AI models. These cases violate the ABCDE checklist, which forms the foundation of melanoma identification and diagnosis in both the clinical sphere and AI model development. Typically, models are trained to identify specific "malignant" features, meaning atypical presentations such as benign-appearing lesions are systematically classified incorrectly.

While the challenge of classifying benign-appearing melanomas is well documented in the current literature (Li et al., 2022), their specific impact on AI systems in diagnosing skin lesions remains unknown. There remains a poor understanding of how lesions that are difficult to classify can be identified, the objective impact of these lesions on the performance of deep AI models, and what mechanisms cause these systematic failures.

5. Paper Overview

This study provides a comprehensive analysis into how convolutional neural networks assess benign-appearing melanomas, and the impact that critical features such as color-uniformity and symmetry influence have on diagnostic results. Quantitative results are combined with interpretability analysis to understand the specific reason models struggle with benign-appearing lesions. This approach moves beyond recognizing the problem with current models, and towards quantifying its risks and impact to support future model improvements.

First, objective metrics for color-uniformity and symmetry are established using the HAM10000 dataset and expert-drawn segmentation masks. Samples will be divided using these metrics into benign-appearing and typical melanomas. The model's performance is then evaluated separately for these two types of representations. Finally, GradCAM analysis is applied to understand the model's attention mechanism in the most challenging cases. The study demonstrates that benign-appearing melanomas do pose significant challenges for AI melanoma detection, though not through the previously assumed mechanisms.

The remainder of the paper is organized as follows: Section 2 reviews related works on melanoma detection using AI technology. Section 3 describes the study's methods for identifying benign-appearing cases, evaluating model performance, and interpreting reasons for failure. Section 4 presents the model's performance results along with attention analysis. Section 5 examines the implications of this experimental study and future directions for research. Finally, section 6 restates key takeaways and suggests improvements for melanoma detection using AI.

II. Related Works

1. AI Use in Dermatology

The true efficacy of AI models at identifying melanoma was first widely recognized in Esteva et al.'s (2017) study which demonstrated a model that attained dermatologist-level classification. The convolutional neural network in this landmark study employed a dataset of 129,450 images for training and was compared to 21 certified dermatologists. The results showed that AI had a promising future in diagnosing skin conditions and expanding access to patients outside the setting of a clinic. Subsequent studies expanded on these findings with increasing success: Brinker et al. (2019) found that a trained CNN model outperformed 136 out of 157 dermatologists, while Pham et al. (2021) found AI performing at a sensitivity and specificity of 85.0% and 95.0% compared to 74.1% and 60.0% for dermatologists.

However, most studies only report aggregate results, meaning performance metrics across lesion subsets such as benign-appearing melanomas is obscured. Detailed analysis of subgroup performance conducted by Combalia et al. (2022) for example reveals concerning disparities between classes, with nevi reporting a 63.6% balanced accuracy while nevi with crust was only correctly classified 34.7% of the time. Other studies such as Marchetti et al. (2023) similarly note that challenging lesions are commonly underrepresented, leading to training bias.

2. AI Failure Modes

Recent research on the failure modes of melanoma classification AI models have also identified other systematic failure modes such as image artifacts, unrelated skin markings, and hair as significant obstacles (Li, 2022). The drastic skin color disparity present in most popular skin lesion datasets also introduces racial bias, as models are unfamiliar with

samples from darker skinned people (Kassem, 2021). These findings suggest that AI models can rely on specific visual distractions instead of the lesion itself, leading to systematic failure that may go unnoticed due to the rarity of cases. In terms of melanoma classification and the approach to benign-appearing lesions, it remains unclear whether models fail due to misdirected attention toward peripheral features or an inability to extract diagnostic information from symmetric, uniformly colored lesions. Simple performance metrics obscure whether the models fail due to focusing on peripheral features instead of the lesion itself, or due to an inability to recognize subtle malignant indicators within benign-appearing lesions.

3. AI Model Interpretability

Facing recurring evidence of systematic failure in melanoma classification by AI models, researchers have developed interpretability methods designed to diagnose the models themselves. Interpretability methods are crucial for improving AI models, as they offer insight into the model's actual failure mechanisms. Researchers can understand which features are prioritized by the model, which are neglected, and whether failures are attention or architecture-based.

Several studies have applied interpretability methods to dermatology AI, though they are most commonly used to validate correct classifications instead of analyzing failure modes. Chanda et al. (2024) analyzed the attention overlap between an AI model and dermatologists and found a 46% overlap when both subjects identified melanoma correctly. Other studies such as Gamage et al. (2024) have also demonstrated that the training and fine tuning process for models improves when aided by interpretability practices. Most studies that delve into model interpretability employ the use of Gradient-weighted Class Activation Mapping or GradCAM, which inspects the convolutional layers of a model and visually reveals the areas of a sample that most influence a prediction. Heatmaps generated by CAM techniques allow model engineers to understand which features influence a model's decision making.

Despite these new insights, studies rarely investigate attention patterns for specific subsets of skin lesions. Matas et al. (2024) used GradCAM to compute the model attention's focused inside lesion masks drawn by clinicians, but the findings were presented as an aggregated mean. Prisacăriu et al. (2024) found with GradCAM implementation that models with single attention mechanisms on global features had the strongest performance and exhibited better understanding of lesion features. However, their findings were presented through mean GradCAMs instead of on a class-by-class basis.

Understanding whether models fail on lesions exhibiting high color-uniformity and symmetry due to misdirected focus or failure to process subtle features is essential for developing future improvements. Results showing models attending appropriately to the skin lesion but unable to reach the correct classification would call for architectural adjustments. On the other hand, attention guidance techniques would be needed to correct for misdirected focus.

4. Similar Challenges for Dermatologists and Research Gap

The systematic failures on benign appearing melanomas of AI models mirror a well-documented challenge in clinical dermatology. Pizzicheta et al. (2004) found that melanotic (lacking pigment) and hypomelanotic (partially pigmented) melanomas are significantly more difficult to diagnose because of their unusual presentations. The study also found that more than

half of the undiagnosed benign-appearing lesions could “masquerade clinically” as benign conditions. Martín-Alcalde et al. (2021) indicates that the challenge extends beyond melanotic and hypomelanotic melanomas, with nevoid melanoma also presenting a common pitfall for dermatologists. Studies such as Di Stefani et al. (2013) have quantified the challenge, finding 42% of melanomas exhibiting benign dermoscopic features. When these features were present, diagnostic accuracy dropped dramatically from 87% to only 54.1%.

The current literature strongly indicates that current AI models are least reliable precisely when the need for assistance is greatest. Moving forward with deep learning models for melanoma detection will require more than incremental improvements on aggregate model performance. A mechanistic understanding of misclassifications, especially lesion types that represent systematic failure modes, is crucial to developing systems that complement human limitations instead of replicating them. This investigation addresses this need by providing an integrated analysis of morphological features, diagnostic performance, and attention mechanisms for malignant melanomas with benign features.

III. Methods

1. Dataset

This study utilized the HAM10000 dataset, which is a large collection of multi-source dermoscopic images of skin lesions. HAM10000 was chosen due to its vast representation of different skin lesion types and open access availability. The dataset contains 10,015 total images from different populations, acquired and stored using various modalities. Each image is accompanied by metadata which includes the histopathology-confirmed or dermatologist-consensus diagnosis, the lesion location, and the patient’s age and sex. While seven categories of pigmented lesions are represented in the dataset, this study focuses on the binary classification of melanomas. Thus, the dataset is split into either melanoma or non-melanoma samples, resulting in 1,113 melanoma cases and 8,902 non-melanoma cases.

2. Segmentation Masks

To accurately quantify lesion characteristics and analyze model attention, this study employed expert-drawn segmentation masks available through the ISIC Archive. These binary masks delineate the exact boundaries of skin lesions, enabling accurate computation of each lesion’s morphological features. The masks were preprocessed by converting PNG images to binary arrays using a set threshold, then individually approved or redrawn by a dermatologist. These masks were chosen for their outstanding accuracy compared to automatically generated masks.

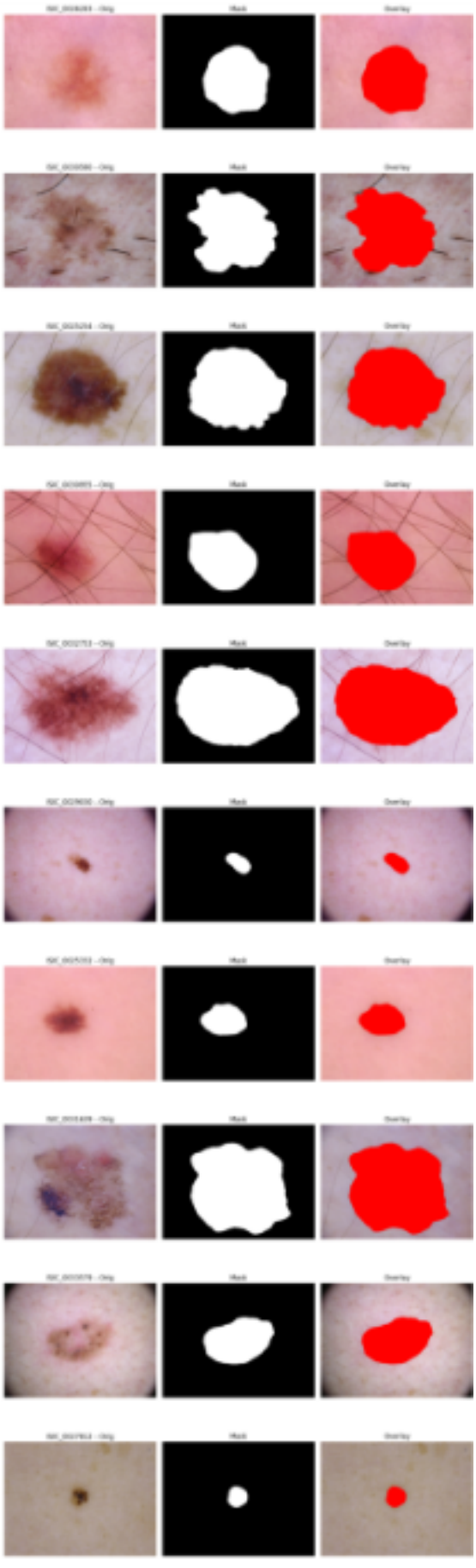

Figure 1: Examples of dermoscopic images and respective segmentation masks for the HAM10000 dataset. Each row includes the original image sample and ID on the left, the binary lesion mask in the middle, and a visual overlay on the right. A variety of lesion morphologies including different sizes, symmetry, and color-uniformity are demonstrated.
Figure generated by student researcher using Python Matplotlib, 2025.

3. Color Uniformity Metrics

Color uniformity was computed using the CIE LAB color space, a color space that provides perceptually uniform color representation. LAB uses three channels, with the L channel expressing lightness, the A channel expressing the green-red component, and the B channel expressing the blue-yellow component. Each image was first converted from RGB into the LAB color space. Then, the standard deviation for each channel was computed. The mean of these three standard deviations was used as the final color uniformity score.
To optimize computational efficiency, images were resized to 256x256 pixels.

4. Symmetry Metrics

Lesion symmetry was assessed through geometric analysis of the segmentation masks. Each lesion mask was first split into equal halves along its central axis. Then, the right half was flipped and overlayed onto the left half for comparison. Next, symmetry was calculated as the Intersection over Union (IoU calculated as the area where pixels match in both halves divided by area occupied by pixels from either half) where a value of 0 indicates no symmetry and a value of 1.0 indicates perfect symmetry.

5. Challenging Case Definition

Benign-appearing melanomas were identified through simultaneous high values in both metrics for uniformity and symmetry computed previously. A grid search was implemented to determine the optimal color-uniformity and symmetry metrics for the dataset. Samples were classified as “challenging” if they had a color uniformity score (mean standard deviation across LAB channels) greater than 8.3089 (74^{th} percentile) and a symmetry score (Intersection over Union) greater than 0.8265 (89^{th} percentile). Using these definitions, 436 benign-appearing cases were identified, making up 4.35% of the dataset.

6. Model Architecture and Training

An EfficientNet-B3 architecture was chosen for its proven performance in medical imaging tasks as well as its computational efficiency compared to other popular models like ResNet-152. The model was initialized with pretrained weights for image recognition from ImageNet, applying transfer learning to lower training time. Finally, the model was adapted for binary classification of melanoma. The AdamW optimizer was chosen due to its improved weight decay, which helps prevent the model from overfitting on the dataset. The loss function used was cross-entropy loss, which is crucial for classification tasks. The learning rate scheduler used was StepLR with a step size of 3 epochs and gamma of 0.1 to encourage fine-tuning later in the training process. 32 images were included in each batch. The model was trained for 5 epochs on Google Colab using a TPU, with image preprocessing conducted on CPU.

For the training process, images were randomly cropped and resized to 224x224 pixels, randomly flipped horizontally and vertically, and augmented with ColorJitter. These randomizations simulate natural variation in data and encourage the model to learn from lesion features instead of the image format. For validation and testing, a center crop was applied to match the size of the image during training. For image normalization, statistics from ImageNet (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) were used to match the pretrained model.

The dataset was divided into three for training, validation, and testing. The three sets were divided using stratified random sampling to maintain class balance, resulting in 7,010 (70% of total) images for training, 1,502 (15% of total) for validation, and 1503 (15% of total) for testing.

7. False Negative Rate Evaluation Metrics

Model performance was evaluated using the False Negative Rate (FNR) on melanoma cases, which is calculated as FNR = FN / (FN + TP), where FN represents melanomas incorrectly classified as benign, and TP represented correctly identified melanomas. The FNR was computed separately for melanoma samples categorized as benign-appearing and those representing typical presentations.

8. Attention Evaluation Metrics

Gradient-weighted Class Activation Mapping (GradCAM) was applied to visualize and quantify model attention patterns. GradCAM was chosen for its interpretability capabilities and its established use in AI medical imaging research. The analysis targeted the final convolutional layer of the EfficientNet-B3 model. For each misclassified melanoma case, GradCAM heatmaps were first generated using the true class as the target. Then, heatmaps were resized to match the original image sizes using bilinear interpolation. Finally, the mean attention of the model concentrated within the lesion boundaries for both benign-appearing and normal melanomas were calculated. The formula is $\Sigma(\text{CAM} \times \text{Mask}) / \Sigma(\text{CAM})$, where CAM represents the attention map generated by GradCAM and Mask represents the binary lesion segmentation mask.

9. Statistical Analysis

The statistical significance of difference between challenging and typical cases was assessed using multiple tests. The two proportion z-test was used to compare false negative rates of benign appearing and normal melanomas, and the independent samples t-test was used to compare attention inside-fraction for benign appearing and normal melanomas. One-sided Fisher's exact test was implemented to test for association between diagnostic accuracy and benign-appearing features. Permutation tests with 5,000 iterations were also implemented for non-parametric validation. 95% bootstrap confidence intervals were computed using 1,000 iterations with replacement. Effect size was measured with Cohen's h for difference in FNR proportions, and Cohen's d for inside fraction difference. P-values from independent tests were combined using Fisher's method when appropriate. Statistical significance was set at $\alpha = 0.05$, with exact p-values reported for transparency.

IV. Results

1. Benign-Appearing Melanoma Categorization

After applying the thresholds for color-uniformity and symmetry (defined as color uniformity ≥ 8.3089 and symmetry ≥ 0.8265) to the HAM10000 dataset, 436 melanomas from the dataset (4.35% of total) were characterized as benign-appearing. Due to the rarity of benign-appearing lesions, the final dataset used for testing only contained 8 benign-appearing representations and 159 typical melanomas.

2. Classification Performance

After training, the EfficientNet-B3 model achieved an overall balanced accuracy of 0.846 for binary melanoma classification on the 1503 image test set. The model

performed markedly differently for the two types of melanomas. The FNR for benign-appearing melanomas and typical melanomas were 87.5% vs 59.7% respectively, resulting in a 27.8 percentage point difference. The significance of this difference was confirmed with the Fisher's exact test ($p = 0.112$) and permutation test ($p = 0.104$), resulting in a combined p-value of 0.065 using Fisher's method. The effect size calculated using Cohen's h was 0.652, indicating a medium-large effect size (0.2 = small, 0.5 = medium, 0.8 = large). The 95% confidence intervals for the false negative rate of benign-appearing and typical melanomas were [62.5%, 100%] and [52.8%, 67.3%] respectively.

3. Attention Analysis

GradCAM analysis was conducted on the 102 misclassified melanoma cases from the test set to understand attention patterns in failure modes. The analysis returned an unexpected result: The mean attention inside-fraction was higher for benign-appearing melanomas than typical cases, at 48.9% and 24.6% respectively. The difference was +24.3% with a p-value of 0.099. A Cohen's d of 1.107 indicated a large effect size (0.2 = small, 0.5 = medium, 0.8 = large).

Metric	Benign-appearing	Typical cases	Difference	P-value	Effect Size
False Negative Rate	87.5% (7/8 samples)	59.7% (95/159 samples)	27.8%	0.065	Cohen's h = 0.652 (medium-large
Mean Attention	48.9% ± 32.8% (n=7)	24.6% ± 21.0% (n=95)	24.3%	0.099	Cohen's d = 1.107 (large

Table 1: Table showing model performance metrics on binary classification of benign-appearing vs. typical melanomas and model attention inside-fraction for the two types of presentations. FNR was demonstrated to be 27.8% higher for benign-appearing melanomas. The mean attention inside-fraction for benign-appearing lesions was greater than typical presentations by 24.3%.

Table created by student researcher, 2025.

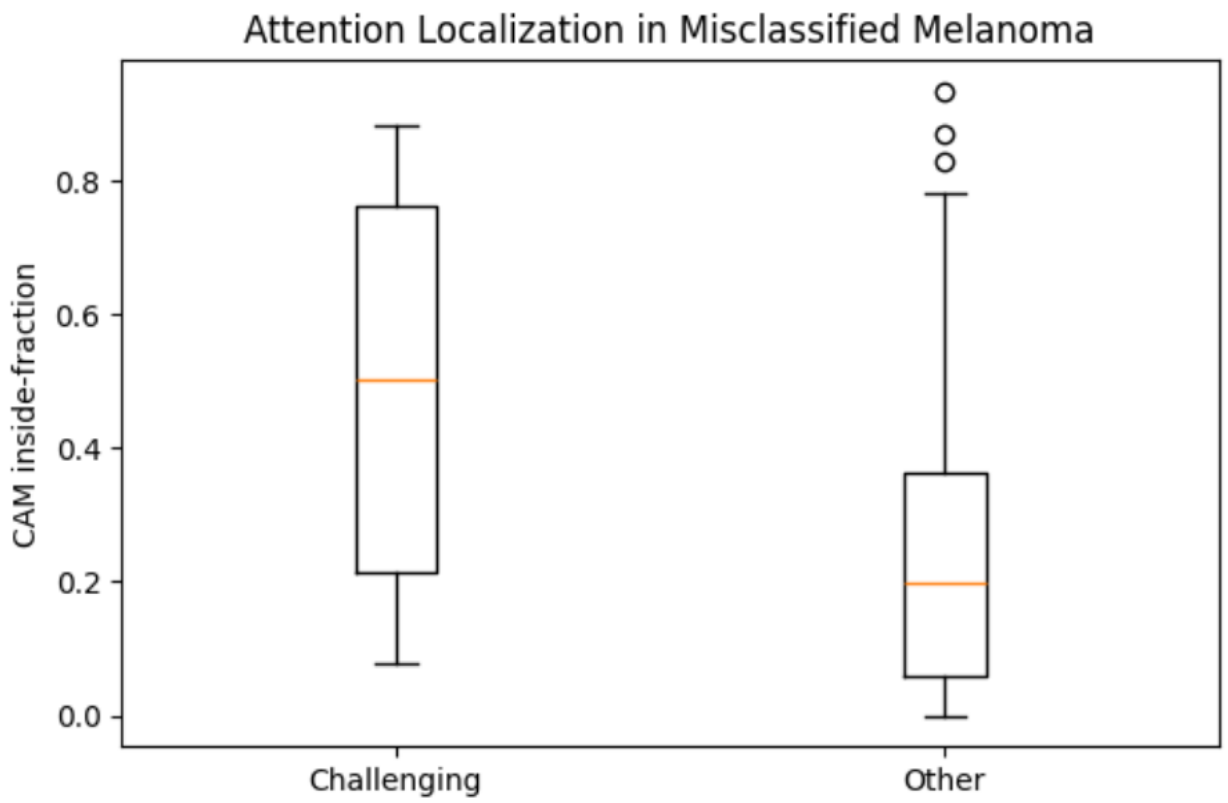

Figure 2: Boxplot of GradCAM analysis of attention inside lesion boundaries for misclassified melanomas. Benign-appearing melanomas showed higher attention inside-fraction (48.9%) compared to typical presentations (24.6%). This contradicts the previous hypothesis that models fail due to distracted attention on benign-appearing samples.
Figure created by student researcher using Python Matplotlib, 2025.

V. Discussion

1. Findings

This study provides the first quantitative analysis of how color-uniformity and symmetry, two features associated with benign presentations, affect deep learning model performance in melanoma identification. The findings confirm that AI models are more likely to fail on lesions exhibiting high color-uniformity and symmetry, with false negative rates being 27.8% higher than typical presentations. However, GradCAM analysis revealed an unexpected explanation for these failures that challenges conventional assumptions about AI attention and typical failure modes.

Contrary to the initial hypothesis that models would fail due to misdirected attention toward peripheral features, the study found that misclassified benign-appearing melanomas actually received more attention within lesion boundaries than typical cases (48.9% vs 24.6%). This suggests that the failure mechanism is not misdirected attention towards image artifacts or irrelevant peripheral features, but rather an inability to recognize subtle features indicating malignancies.

2. Explanations For Failure

The paradoxical finding that models have appropriately adjusted their attention to skin lesions yet fail diagnostically points to limitations in the current deep learning architecture used for training. There are several non-mutually exclusive mechanisms that may explain this observation.

One possible explanation is a training distribution mismatch. There is a severe underrepresentation of benign-appearing cases (4.35% of total dataset), likely contributing to poor generalization. The disparity in the amount of benign-appearing and typical samples encourages the model to fit itself for obvious malignant features and develop decision boundaries that inadequately separate benign-appearing melanomas from benign lesions. This can all occur despite attention mechanisms being appropriate.

Another possible explanation is that the model is not optimized to extract subtle features from dermoscopic imaging. In addition to symmetry and color-uniformity, models need to extract minute morphological and textural features to distinguish between benign-appearing melanomas from benign lesions. This suggests that current CNN architectures, which are optimized for recognizing relatively general objects and shapes, may lack the depth needed for fine-grained medical image classification.

3. Comparison to Human Performance

These results have profound implications for the integration of AI in clinical melanoma diagnosis. Rather than complementing human shortcomings, AI models tend to replicate the mistakes that dermatologists make. Using the ABCDE checklist and other common approaches, melanomas can be easily missed on a first pass. Standard approaches place an emphasis on reliably identifying melanomas that have clear malignant indicators, with less focus on lesions with benign features. While AI has been proven to enhance diagnostic accuracy and diagnostic confidence when used with dermatologists, the current results show that the technology fails precisely where assistance is most crucial.

4. Limitations

The most significant limitation of this study stems from the rarity of benign-appearing melanomas in the training data, resulting in only eight such cases in our test set. Although the observed effect sizes were large (Cohen's $h = 0.652$ and Cohen's $d = 1.107$), the p value just shy of significance and wide confidence intervals reflect substantial statistical uncertainty.

The study also exclusively uses the HAM10000 dataset, which introduces further constraints for the generalizability of the findings. The 2D images included in the dataset may not capture the full spectrum of diagnostic information available to clinicians in practice, including detailed patient metadata and three-dimensional features. Additionally, the geographic and demographic composition of the HAM10000 dataset may not represent the full diversity of the global population, a well-documented flaw that has been proven to cause issues with AI performance across different skin tones.

Finally, the definition of "benign appearing" melanomas relied on specific threshold values for color uniformity and symmetry, which may not fully capture the criteria used by dermatologists when identifying diagnostically challenging cases. The use of the CIE LAB color space and IoU for assessing color uniformity and symmetry, though quantitative and

theoretically justifiable, represents one approach among several possible strategies to distinguish benign-appearing melanomas.

5. Future Direction

The unexpected finding that models fail despite appropriate spatial attention suggests that fundamental architectural innovations will be necessary to address the challenge of benign-appearing melanomas. Vision transformer architectures, which have a greater capacity for multi-scale attention and flexibility, may be better suited for capturing subtle diagnostic patterns that benign-appearing melanomas exhibit. Recent advances made by Cirrincione et al. (2023) for example have shown compelling evidence of transformer models achieving superior performance when compared to state-of-the-art techniques. Additionally, Hamsalekha et al. (2024) showed that hybrid CNN-transformer models can leverage the complementary strengths of the two architectures to produce performance suitable for clinical integration. These architectural innovations provide a clear pathway for overcoming the feature extraction inadequacies of current CNN models. Future work should systematically evaluate whether specialized architectures like vision transformers can reduce the dramatic false negative rate disparity between benign-appearing and typical melanoma cases.

Furthermore, the severe class imbalance between benign-appearing and typical melanomas necessitates new training strategies that force models to draw more accurate decision boundaries. Changes to the focal loss function specifically geared towards extremely rare classes, combined with sampling strategies that ensure adequate representation of benign-appearing melanomas could improve model performance. Curriculum learning strategies that progressively expose the models to increasing levels of diagnostic difficulty may further improve the model's ability to attend to benign-appearing melanomas.

The findings from model attention analysis further point to the potential value of incorporating multi-modal imaging in model training. Techniques such as reflectance confocal microscopy, optical coherence tomography, or sequential clinical photography could be combined to provide previously unavailable spatial and temporal information necessary to diagnosing melanoma. Beyond improvements to imaging modalities, the incorporation of patient metadata including demographics, family history, and anatomical location could help computational models diagnose skin lesions holistically, similar to dermatologists.

The safety implications of the current findings cannot be overlooked in future research. Given that benign-appearing melanomas pose challenges for both human experts and artificial intelligence systems, future research should prioritize developing models with interpretability in mind to ensure effective communication of the model's reasoning. Research should additionally focus on implementing robust monitoring systems for deployed models and clear clinical guidelines that account for known failures, ensuring that the technology is integrated safely and effectively in clinical settings.

VI. Conclusion

This study provides the first quantitative evidence that benign-appearing melanomas represent a critical blind spot in AI melanoma detection. These melanomas exhibiting high color uniformity and symmetry demonstrated an 87.5% false negative rate compared to 59.7% for typical presentations. The drastic difference in diagnosis

performance could translate directly into patient wellbeing, given melanoma's dramatic survival rate from near 100% to 34.6% when metastasized (SEER, 2025).

Most surprisingly, GradCAM analysis revealed that models fail to detect benign-appearing melanomas despite appropriate focus within lesion boundaries (48.9% for benign-appearing vs 24.6% for typical). This contradicts assumptions about misdirected focus causing systematic failure and instead indicates that current CNN architecture may lack the capability to extract subtle diagnostic features from benign-appearing lesions.

These results carry immediate clinical implications, as AI systems appear to replicate human shortcomings instead of complementing them. Moving forward requires research exploring fundamental CNN architecture innovations, vision transformer and hybrid CNN architectures, specialized training strategies for rare lesion subsets, and clinical protocols that acknowledge systematic failures.

By quantifying this previously unexplored failure mode, this study establishes that addressing benign-appearing melanomas is essential for developing AI tools that truly aid in the clinical process and reduce melanoma's devastating impact.

Bibliography

American Cancer Society. (2025). *Melanoma skin cancer statistics*. Melanoma Skin Cancer Statistics | American Cancer Society. https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html

Brinker, T. J., Hekler, A., Enk, A. H., Berking, C., Haferkamp, S., Hauschild, A., Weichenthal, M., Klode, J., Schadendorf, D., Holland-Letz, T., von Kalle, C., Fröhling, S., Schilling, B., & Utikal, J. S. (2019). Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, *119*, 11–17. https://doi.org/10.1016/j.ejca.2019.05.023

Chanda, T., Hauser, K., Hobelsberger, S., Bucher, T.-C., Garcia, C. N., Wies, C., Kittler, H., Tschandl, P., Navarrete-Dechent, C., Podlipnik, S., Chousakos, E., Crnaric, I., Majstorovic, J., Alhajwan, L., Foreman, T., Peternel, S., Sarap, S., Özdemir, İ., Barnhill, R. L., … Brinker, T. J. (2024). Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-023-43095-4

Chen, J. Y., Fernandez, K., Fadadu, R. P., Reddy, R., Kim, M.-O., Tan, J., & Wei, M. L. (2025). Skin cancer diagnosis by lesion, physician, and examination type. *JAMA Dermatology*, *161*(2), 135. https://doi.org/10.1001/jamadermatol.2024.4382

Chen, J., Stanley, R. J., Moss, R. H., & Van Stoecker, W. (2003). Colour analysis of skin lesion regions for melanoma discrimination in clinical images. *Skin Research and Technology*, *9*(2), 94–104. https://doi.org/10.1034/j.1600-0846.2003.00024.x

Cirrincione, G., Cannata, S., Cicceri, G., Prinzi, F., Currieri, T., Lovino, M., Militello, C., Pasero, E., & Vitabile, S. (2023). Transformer-based approach to melanoma detection. *Sensors*, *23*(12), 5677. https://doi.org/10.3390/s23125677

Combalia, M., Codella, N., Rotemberg, V., Carrera, C., Dusza, S., Gutman, D., Helba, B., Kittler, H., Kurtansky, N. R., Liopyris, K., Marchetti, M. A., Podlipnik, S., Puig, S., Rinner, C., Tschandl, P., Weber, J., Halpern, A., & Malvehy, J. (2022). Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: The 2019 International Skin Imaging Collaboration Grand Challenge. *The Lancet Digital Health*, *4*(5). https://doi.org/10.1016/s2589-7500(22)00021-8

Debelee, T. G. (2023). Skin lesion classification and detection using Machine Learning Techniques: A systematic review. *Diagnostics*, *13*(19), 3147. https://doi.org/10.3390/diagnostics13193147

Di Stefani, A., Massone, C., Soyer, H. P., Zalaudek, I., Argenziano, G., Arzberger, E., Lozzi, G. P., Chimenti, S., & Hofmann-Wellenhof, R. (2013). Benign dermoscopic features in melanoma. *Journal of the European Academy of Dermatology and Venereology*, *28*(6), 799–804. https://doi.org/10.1111/jdv.12192

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with Deep Neural Networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

European Medical Journal. (2024, November 18). *Experience And Dermoscopy Improve Skin Cancer Diagnosis Accuracy*. https://www.emjreviews.com/en-us/amj/dermatology/news/experience-and-dermoscopy-improve-skin-cancer-diagnosis-accuracy/

Gamage, L., Isuranga, U., Meedeniya, D., De Silva, S., & Yogarajah, P. (2024). Melanoma skin cancer identification with explainability utilizing mask guided technique. *Electronics*, *13*(4), 680. https://doi.org/10.3390/electronics13040680

Haddadin, S., & Ganti, L. (2024). The use of artificial intelligence to detect malignant skin lesions. *Mayo Clinic Proceedings: Digital Health*, *2*(2), 241–245. https://doi.org/10.1016/j.mcpdig.2024.04.003

Hamsalekha, Glan Devadas, & Satheesha. (2025). A novel deep learning approach for automated melanoma classification using hybrid CN N and Vision Transformer model. *Fusion: Practice and Applications*, *19*(2), 92–101. https://doi.org/10.54216/fpa.190207

Jiang, C., Jain, N. P., & Stewart, C. L. (2025). Amelanotic melanoma: Clinical presentation, diagnosis, and Management. *Clinics in Dermatology*, *43*(1), 10–15. https://doi.org/10.1016/j.clindermatol.2025.01.009

Kassem, M. A., Hosny, K. M., Damaševičius, R., & Eltoukhy, M. M. (2021). Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. *Diagnostics*, *11*(8), 1390. https://doi.org/10.3390/diagnostics11081390

Li, Z., Koban, K. C., Schenck, T. L., Giunta, R. E., Li, Q., & Sun, Y. (2022a). Artificial Intelligence in dermatology image analysis: Current developments and future trends. *Journal of Clinical Medicine*, *11*(22), 6826. https://doi.org/10.3390/jcm11226826

Li, Z., Koban, K. C., Schenck, T. L., Giunta, R. E., Li, Q., & Sun, Y. (2022b). Artificial Intelligence in dermatology image analysis: Current developments and future trends. *Journal of Clinical Medicine*, *11*(22), 6826. https://doi.org/10.3390/jcm11226826

Li, Z., Koban, K. C., Schenck, T. L., Giunta, R. E., Li, Q., & Sun, Y. (2022c). Artificial Intelligence in dermatology image analysis: Current developments and future trends. *Journal of Clinical Medicine*, *11*(22), 6826. https://doi.org/10.3390/jcm11226826

Marchetti, M. A., Cowen, E. A., Kurtansky, N. R., Weber, J., Dauscher, M., DeFazio, J., Deng, L., Dusza, S. W., Haliasos, H., Halpern, A. C., Hosein, S., Nazir, Z. H., Marghoob, A. A., Quigley, E. A., Salvador, T., & Rotemberg, V. M. (2023). Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (prove-AI study). *Npj Digital Medicine*, *6*(1). https://doi.org/10.1038/s41746-023-00872-1

Martín-Alcalde, J., Gamo-Villegas, R., Floristán-Muruzábal, M. U., Pampín-Franco, A., Pinedo-Moraleda, F., & López-Estebaranz, J. L. (2021a). Nevoid melanoma: Dermoscopic and in vivo reflectance confocal microscopic aspects in 4 cases. *JAAD Case Reports*, *11*, 132–136. https://doi.org/10.1016/j.jdcr.2021.03.044

Martín-Alcalde, J., Gamo-Villegas, R., Floristán-Muruzábal, M. U., Pampín-Franco, A., Pinedo-Moraleda, F., & López-Estebaranz, J. L. (2021b). Nevoid melanoma: Dermoscopic and in vivo reflectance confocal microscopic aspects in 4 cases. *JAAD Case Reports*, *11*, 132–136. https://doi.org/10.1016/j.jdcr.2021.03.044

Matas, I., Serrano, C., Silva, F., & Serrano, A. (2024). *AI-Driven Skin Cancer Diagnosis: Grad-CAM and Expert Annotations for Enhanced Interpretability*. https://doi.org/10.48550/arXiv.2407.00104

Naeem, A., Farooq, M. S., Khelifi, A., & Abid, A. (2020). Malignant melanoma classification using Deep Learning: Datasets, Performance Measurements, challenges and opportunities. *IEEE Access*, *8*, 110575–110597. https://doi.org/10.1109/access.2020.3001507

Pham, T.-C., Luong, C.-M., Hoang, V.-D., & Doucet, A. (2021a). Ai outperformed every dermatologist in dermoscopic melanoma diagnosis, using an opti mized deep-CNN architecture with custom mini-batch logic and loss function. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-96707-8

Pham, T.-C., Luong, C.-M., Hoang, V.-D., & Doucet, A. (2021b). Ai outperformed every dermatologist in dermoscopic melanoma diagnosis, using an opti mized deep-CNN architecture with custom mini-batch logic and loss function. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-96707-8

Pizzichetta, M. A., Talamini, R., Stanganelli, I., Puddu, P., Bono, R., Argenziano, G., Veronesi, A., Trevisan, G., Rabinovitz, H., & Soyer, H. P. (2004). Amelanotic/hypomelanotic melanoma: Clinical and dermoscopic features. *British Journal of Dermatology*, *150*(6), 1117–1124. https://doi.org/10.1111/j.1365-2133.2004.05928.x

Prisăcariu, A.-M., & Ferariu, L. (2024). Skin cancer diagnosis using CNN with attention mechanisms based on grad-cam. *2024 28th International Conference on System Theory, Control and Computing (ICSTCC)*, 107–112. https://doi.org/10.1109/icstcc62912.2024.10744754

SEER. (2025). *Common cancer sites - cancer stat facts*. https://seer.cancer.gov/statfacts/html/common.html

Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., & Jemal, A. (2025). Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, *75*(1), 10–45. https://doi.org/10.3322/caac.21871

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human–computer collaboration for Skin cancer recognition. *Nature Medicine*, *26*(8), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0