CS610 Applied Machine Learning
Project Report

Skin Cancer Detection

Group 5

Tran Binh Minh**, 01323752**
Shwe Tin Aung, **01439396**
Aditya Vijay, **01522128**
Liu Chih Yuan, **01500507**
Lope Immanuell IV Cudiamat Santos, **01520025**

Repository
https://github.com/AdityaVijay1/AML_Group5

Models
https://drive.google.com/drive/folders/1rrRTIrI0sHaR4334fQK8tim-0rIQSgT4?usp=sharing

6th July 2025

I. Introduction

Melanoma is among the most aggressive forms of skin cancer, with survival rates highly contingent upon the timeliness and accuracy of its detection. Early-stage melanoma detection significantly improves survival prospects, boasting nearly a 100% five-year survival rate. However, this rate drastically declines once the disease advances, highlighting the critical need for reliable diagnostic tools and expertise.

There are current diagnostic methods available, yet each has their own advantages and disadvantages. Visual inspection or using the naked eye is simple and widely used but highly subjective and prone to human error. The gold standard, biopsy, is accurate but rather invasive, time-consuming, and makes patients uncomfortable. Non-contact testing is also available; one example is Reflectance Confocal Microscopy which provides high-resolution diagnostics but expensive and has limited availability.

This project aims to address these gaps by developing a deep learning-based diagnostic system focused on classifying skin lesion images as benign or malignant. The primary objectives of the project are to ensure reliable predictions across a diverse dataset, to accurately identify malignant cases to minimize missed diagnoses, and to provide clear explanations to clinicians for each prediction made by the deep learning model.

Like most real-world datasets, this project would also need to consider and address severe class imbalance, missing metadata, skin tone biases, and inconsistent image quality. By prioritizing explainability, the developed model ensures clinical decision-making needs are met through transparent and interpretable AI predictions. Ultimately, this project seeks to enhance clinician trust, improve diagnostic efficiency, and provide a more accessible, reliable, and scalable solution for melanoma detection.

II. Literature Review

Recent advances in deep learning have significantly impacted automated skin cancer detection, with convolutional neural networks (CNNs), especially EfficientNet, playing a prominent role. EfficientNet models, leveraging compound scaling for improved efficiency, have consistently shown high accuracy and reduced computational requirements compared to earlier CNN architectures (Gessert et al., 2021). Efficientnet-B7, for instance, achieved 84.4% accuracy on multi-class skin lesion datasets, while being substantially smaller and faster than other CNN models (Gessert et al., 2021). Ensembles incorporating EfficientNet further demonstrated strong performance, highlighting its suitability for sensitive melanoma detection (Haenssle et al., 2020).

Attention mechanisms integrated into EfficientNet have enhanced model interpretability, crucial for clinical adoption, by emphasizing image regions as the most critical to predictions (Nawaz et al., 2023). Despite these advancements, interpretability remains an active research challenge, as clear and transparent explanations are essential to clinical trust (Nawaz et al., 2023).

Vision Transformers (ViTs) have also recently emerged in skin lesion classification, addressing CNN limitations by capturing global context through self-attention mechanisms. Studies indicate that ViT-based models, including CNN-ViT hybrids, achieve comparable or superior accuracy to traditional CNN models (Li et al., 2021). However, ViTs typically require extensive training data, presenting practical limitations in dermatology applications, particularly when datasets are limited (Khan et al., 2023).

Challenges persist, notably dataset imbalance and limited generalization. Class imbalance, a significant obstacle, often leads to biases and misdiagnoses. Researchers have explored augmented data generation and

customized loss functions to counteract this imbalance (Haenssle et al., 2020). Domain shifts also pose generalization challenges when deploying models trained on dermoscopic images across diverse populations or image sources. Techniques such as transfer learning and domain adaptation are actively being explored to enhance model robustness (Gessert et al., 2021).

Our project aligns closely with these insights, utilizing EfficientNet for its proven balance of accuracy, sensitivity, and interpretability. By addressing dataset imbalance and generalization, our solution aims for robust, clinically reliable melanoma detection. Alongside this for further model experimentation and testing we have used the EfficientNet-B0 model, due to computational constraints under the assumption that the better models (B1-B7) would perform even better.

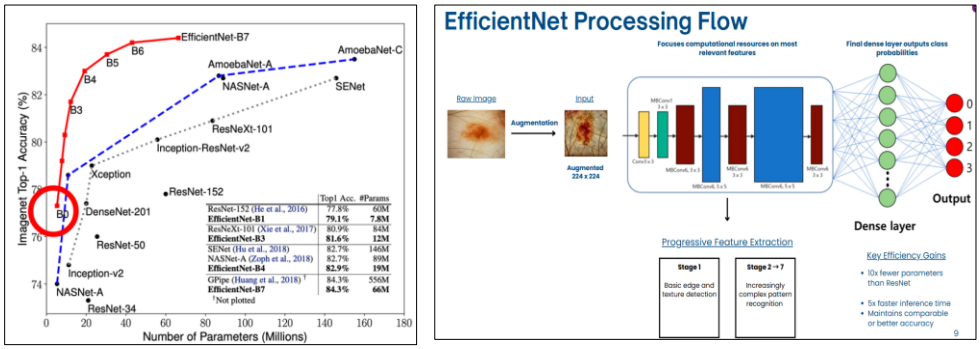


Figure 1. Overview of pretrained models and its respective performance.

III. Dataset Overview

The dataset selected for this project is the SIIM-ISIC 2020 Melanoma Classification dataset, provided by the International Skin Imaging Collaboration (ISIC, 2020). Dataset consists of 33,126 dermoscopic images obtained from a total of 2,056 patients. These images originated from multiple institutions, including centers in Barcelona, Vienna, Sloan Kettering (USA), Queensland (Australia), and Athens.

Images in the dataset were categorized based on clinical diagnosis, with a severely imbalanced class distribution, where 584 images are labeled as melanoma (malignant) while the remaining 32,542 images were class benign. The class imbalance poses a significant challenge for training a robust melanoma classification model; image pre-processing strategies must be implemented to enhance model reliability and performance.

IV. Data Augmentation

Initially, all images were resized to dimensions of 224 x 224 pixels, as this is the optimal input size for Efficientnet-B0, a CNN (convolution neural network) architecture which expects the input of this resolution to achieve balance between computational efficiency and classification accuracy (Tan & Le, 2019).

Subsequently, geometric transformations including transposition and random vertical and horizontal flips were applied to the images. Transpose switches the x axis and y axis of the photo while vertical and horizontal flips are creating mirrored images. These augmentations were implemented to introduce variability in image orientation and spatial positioning, reducing the model's reliance on fixed layouts (Buslaev et al., 2018).

Photometric augmentations were employed to simulate real-world variations in lighting conditions and skin tone diversity. Specifically, RandomBrightnessContrast was used to randomly adjust image brightness and contrast,

replicating variations in illumination, while HueSaturationValue were applied to modify the hue, saturation, and brightness components of the images, simulating realistic variations in skin coloration and imaging equipment inconsistencies. These transformations were set to increase the variability of skin tones, thereby enhancing the model's ability to perform well in different skin tone cases (Buslaev et al., 2018).

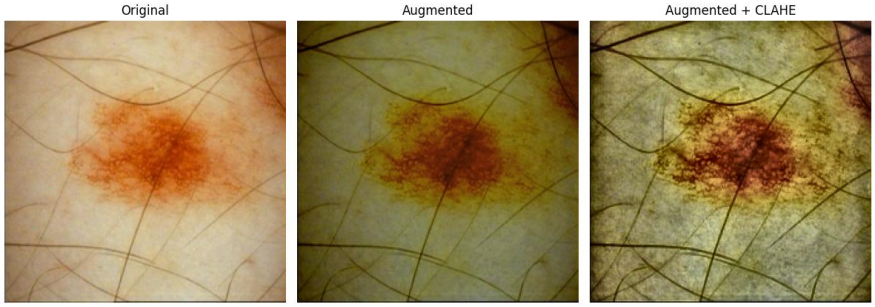


Figure 2. Sample images from original, augmented, and augmented with CLAHE.

Contrast-Limited Adaptive Histogram Equalization (CLAHE) was utilized to enhance local contrast within the images. CLAHE operates on a tile-by-tile basis; its main functions are to enhance the subtle features within shadowed or overly brightened areas, thus highlighting melanoma. This works especially well in images of darker skin tones as there is lower contrast between melanoma and the skin. Importantly, CLAHE's clip limit parameter also prevents excessive noise amplified when increasing contrast, maintaining image quality (Zuiderveld, 1994).

V. Model Training and Experimentation

5.1 Model Selection

In this evaluation, we compared two of the most advanced deep learning architectures for skin cancer detection: the Vision Transformer (ViTI16) and EfficientNet. These models were selected due to their exellent performance in medical image classification tasks. The results clearly show that the Vision Transformer outperforms EfficientNet across all key evaluation metrics. Specifically, ViT achieved higher accuracy (0.75 vs. 0.72), precision (0.6744 vs. 0.65), recall (0.725 vs. 0.65), and AUC (0.8342 vs. 0.814), along with a significantly lower loss (0.9981 vs. 1.6058). Among these, the improvement in recall is particularly critical in the context of malignant skin cancer detection, where false negatives must be minimized to ensure early and accurate diagnosis.

Visual Transformer

Metric	Result
Loss	0.9981
Accuracy	0.75
Precision	0.6744
Recall	0.725
AUC	0.8342

EfficientNet

Metric	Result
Loss	1.6058
Accuracy	0.72
Precision	0.65
Recall	0.65
AUC	0.814

Table 1. Initial results from Visual Transformer and EfficientNet models.

While these results are consistent with existing literature showing that Vision Transformers often outperform CNN-based models in various vision tasks (e.g., Dosovitskiy et al., 2020; Liu et al., 2021), our decision to adopt EfficientNet for downstream use was guided by clinical needs. Despite slightly lower performance, EfficientNet provides far superior explainability via Grad-CAM visualizations. As noted in the study, radiologists expressed a clear preference for the CNN-based model's heatmaps, which more intuitively highlighted affected skin regions. This enhanced interpretability is essential for clinical trust and validation, making EfficientNet the more

appropriate choice for integration into real-world medical workflows where explainability is as important as accuracy.

5.2 Full data training (33k) for Original and Resize image

The initial phase of experimentation focused on evaluating the trade-off between computational resources and model performance by comparing training on original-sized images versus resized datasets. This investigation was prompted by the substantial computational requirements associated with processing high-resolution medical imagery and the need to establish efficient training protocols for resource-constrained environments as well as the norm of resizing images that directly fit what the CNN model will use (Talebi & Milanfar, 2021).

Two parallel training sessions were conducted using identical Efficientnet-B0 architectures, with the primary distinction being input image dimensions. The original dataset maintained native image resolutions, while the resized standardized all images to optimized dimensions. Both experiments were limited to a single epoch to establish baseline performance metrics and computational benchmarks.

The training on original-sized images yielded an accuracy of 98.25% with an Area Under the Curve (AUC) of 84.75%. However, the recall metric remained at 0, indicating the model's inability to correctly identify positive cases within the test set. The computational overhead was substantial, requiring 4 hours of CPU processing time, which was reduced to 1 hour and 30 minutes when utilizing A100 GPU acceleration. The resized dataset experiment demonstrated comparable accuracy (98.25%) but with a slightly reduced AUC of 80.61%. The recall metric remained problematic at 0, consistent with the original-sized image results. Significantly, computational efficiency improved markedly, with CPU training time reduced to 2 hours and 30 minutes, and GPU training time decreased to 40 minutes.

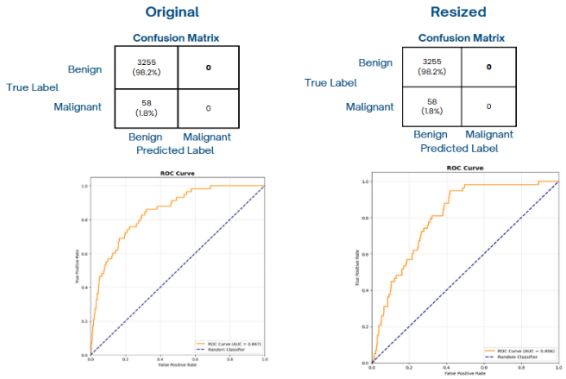


Figure 3. Initial model performance.

The analysis showed no statistical significance in performance between the original and resized datasets. This supports image resizing as a practical strategy for faster training, reducing GPU time by 33% and CPU time by 37.5%—without sacrificing accuracy. The observation of improved AUC performance in preliminary extended training sessions prompted an experiment into optimal epoch ranges. This phase aimed to determine whether additional training iterations could address the concerning recall deficit observed in initial experiments. Training for two epochs on the resized image dataset improved AUC from 73.97% to 83.69% while maintaining 98.25% accuracy, demonstrating significant benefits from extended learning.

Due to computational constraints and the need for systematic comparison across multiple preprocessing approaches, the dataset was strategically reduced to 1,000 images from the original 33,125 samples. Three distinct dataset variants were prepared: original images maintaining native characteristics, resized images

standardized to 224×224×3 dimensions, and augmented images enhanced through Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm implementation.

Training on original images within the reduced dataset yielded an accuracy of 73%, AUC of 80.79%, and notably, a recall of 60%. This marked the first instance of meaningful recall performance, suggesting that dataset size reduction, while limiting overall learning potential, enabled more balanced class representation during training. The resized dataset variant achieved superior accuracy at 87.5% but demonstrated reduced AUC performance at 64% and recall at 50%. This pattern indicates potential overfitting to the majority class, with improved overall accuracy masking reduced sensitivity to minority class detection.

The CLAHE-augmented dataset produced intermediate results with 83.5% accuracy, 72.11% AUC, and 60% recall. The contrast enhancement appeared to improve feature extraction capabilities while maintaining a reasonable balance across performance metrics. Through systematic evaluation across 15 epochs, the optimal training duration was identified within the 10-12 epoch range. This finding provides crucial guidance for full-scale model training, suggesting that extended training beyond this range may result in diminishing returns or potential overfitting.

5.3 Class Imbalance Resolution

The third experimental phase specifically addressed the critical challenge of class imbalance inherent in medical datasets, where malignant cases typically represent a small fraction of total samples. Given the clinical significance of false negative predictions in cancer detection, improving recall performance became the primary objective.

The original dataset exhibited severe class imbalance with only 500 malignant cases among 33,125 total images, representing merely 1.51% of the dataset. This extreme imbalance explained the consistently poor recall performance observed in previous experiments, as the model learned to optimize the majority of the class. Three approaches were conducted as part of improving the performance due to the class imbalance. The initial sampling (985:15 ratio), which is closely like the actual ratio, yielded no meaningful results, proving the need for a better split. The ratio was then increased to 950:50 split, which provided an insufficient improvement on recall, suggesting a need for more aggressive class rebalancing. An aggressive 600:400 split was the last approach, which yielded substantial improvement to key metrics with 70% accuracy, 79.75% AUC, and 80% recall.

The dramatic improvement in recall performance through aggressive class rebalancing demonstrates the potential for artificial augmentation strategies in addressing medical dataset limitations. Indicating tests on generalization of skin tone for classification and most importantly how the model would perform on images naturally taken images of parts of the body that do not have the same focus on the lesion like the original data set (reference in the appendix the two images from ISIC Dataset and Stanford) showing potential overfitting to well-focused lesion images.

5.4 Skin tone Imbalance resolution

The foundational training and evaluation of our EfficientNet-B0 model utilized the ISIC 2020 Challenge dataset which, while comprehensive in scope, presents significant limitations in demographic representation and imaging conditions. The dataset predominantly contains images of lesions on individuals with pale or white skin tones, captured under controlled clinical conditions with optimal lighting, positioning, and focus on the specific area of interest. These standardized imaging conditions, while beneficial for initial model development, create a substantial gap between training data characteristics and real-world deployment scenarios.

The homogeneous representation in the ISIC dataset raises critical concerns regarding model performance across diverse populations, particularly given documented disparities in dermatological AI performance across varying skin tones. Additionally, the controlled imaging environment fails to capture the variability inherent in real-world diagnostic scenarios, where images may be captured using consumer devices under suboptimal lighting conditions. To address these limitations and evaluate model generalization capabilities, we conducted comprehensive testing using an independent Stanford dataset that closely represents real-world usage scenarios. The Stanford dataset contains images captured in naturalistic settings using consumer-grade devices such as smartphones, reflecting the imaging conditions likely to be encountered in telemedicine applications, self-screening scenarios, and point-of-care diagnostics in resource-limited settings.

Data augmentation was implemented specifically designed to increase skin tone diversity of the training dataset. Through computational image processing techniques, synthetic variations of existing training images that simulated darker skin tones were generated while preserving lesion characteristics and morphological features. This augmentation approach expanded the training dataset to 1,200 benign and 800 malignant cases, incorporating representations across white, medium, and dark skin tone categories.

The demographically augmented model demonstrated substantial improvement in generalization performance, achieving a recall of 59.65% when evaluated against the Stanford dataset. This 31% improvement in recall performance provides compelling evidence that explicit demographic diversification in training data significantly enhances model robustness across population groups. The improvement was observed despite maintaining the same testing dataset, indicating that exposure to diverse skin tone representations during training improved the model's ability to extract universal features of malignant lesions independent of demographic characteristics. Additionally, it is important to note that the Stanford dataset was purely used for testing, with no data leakage for the training dataset. The success of computational skin tone augmentation suggests a viable approach for addressing demographic bias in medical AI systems, particularly in scenarios where diverse training data may be limited due to historical healthcare access disparities or clinical trial patterns.

Significantly, the Stanford dataset images exhibit greater variability in terms of anatomical locations, with lesions documented across diverse body regions rather than the standardized presentations typical in clinical photography. The images also demonstrate varying degrees of focus, lighting conditions, and skin surface angles, providing a more rigorous test of model robustness and practical applicability. Importantly, none of the Stanford dataset images were included in the training process, ensuring that performance evaluation represents true generalization rather than memorization of training examples.

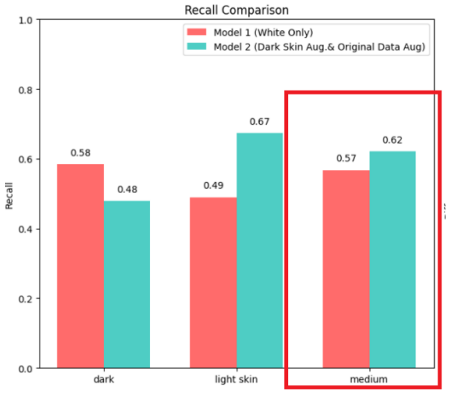


Figure 4. Performance across two different models, 3 types of skin tone.

Model 2, which includes both dark skin augmentation and original data, outperforms Model 1 on light and medium skin tones-rising from 0.49 to 0.67 and from 0.57 to 0.62, respectively. This suggests that incorporating darker skin tone data not only enhances fairness but also improves the model’s generalization across diverse skin

tones. However, it's worth mentioning that despite this augmentation, recall on dark skin remains the lowest among all groups. This indicates room for improvement and raises potential concerns such as limitations in the quality of the Stanford test dataset, challenges in accurately simulating darker skin textures during augmentation, or the need for even more representative samples. Further investigation in later research is necessary to close this performance gap.

VI. Model Explainability

In our skin cancer detection project, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) to interpret predictions made by our EfficientNet-based CNN model. Grad-CAM generates heatmaps to highlight regions of the image crucial for model predictions. This is essential in medical applications, ensuring the model focuses specifically on lesions rather than unrelated image areas. Technically, Grad-CAM calculates gradients of the target class score with respect to the final convolutional feature maps of EfficientNet, identifying the importance of each feature in predictions (Selvaraju et al., 2017).

We implemented Grad-CAM in PyTorch, specifically targeting the last convolutional layer of EfficientNet, responsible for capturing high-level image features. By employing forward and backward hooks, we captured feature maps and corresponding gradients without altering the model architecture. In the forward pass, we extracted activation maps from this targeted convolutional layer. During the backward pass, we computed gradients for the chosen class, providing insights into how sensitive the class score was to individual feature map activations. The gradients were spatially averaged to derive channel-wise importance weights, which we then applied to the feature maps. After combining these weighted feature maps, we generated a final heatmap highlighting image regions positively influencing the predicted class. This heatmap was then resized and superimposed on the original image for intuitive visualization.

Figure 5. Grad-CAM explainability on correctly identified malignant image.

Grad-CAM proved valuable in assessing our model's performance across diverse skin tones. By examining heat maps generated from combined datasets (standard and dark-skin augmented images), we confirmed improved model focus and fairness. Although inconsistencies remained in certain cases, Grad-CAM significantly enhanced our model's interpretability, ensuring its clinical reliability. Overall, integrating Grad-CAM was extremely useful for qualitatively evaluating model behavior across diverse skin tones, allowing us to ensure that the model's attention remained on clinically relevant features (the lesion) regardless of skin background.

VII. Conclusion and Recommendations

This comprehensive experimentation framework has revealed critical insights into the development and deployment of Efficientnet-B0 for skin cancer classification. The computational efficiency gains achieved through image resizing, the importance of extended training periods for minority class detection, the necessity of aggressive class rebalancing for meaningful recall performance, the integration of explainable AI for clinical acceptance, and the critical importance of demographic diversity in training data collectively inform a roadmap for responsible and effective medical AI development. These findings emphasize that successful deployment of AI diagnostic tools requires careful consideration of computational constraints, clinical requirements, and health equity principles throughout the development process.

Future research should prioritize the collection and integration of more demographically diverse datasets, particularly with respect to skin tone, age, and anatomical site. Collaborations with international medical institutions and open data initiatives can help address current limitations in representation. Furthermore, while synthetic augmentation of skin tones has shown promise, further exploration of advanced augmentation techniques—including generative adversarial networks (GANs) for realistic image synthesis—may enhance model robustness and generalizability. Lastly, further development and integration of explainability frameworks beyond Grad-CAM, such as SHAP or LIME, may provide deeper insights into model decision-making and foster greater clinician trust.

References

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). *Swin Transformer: Hierarchical vision transformer using shifted windows*. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10012–10022. https://arxiv.org/abs/2103.14030

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-CAM: Visual explanations from deep networks via gradient-based localization*. *International Journal of Computer Vision, 128*(2), 336-359. https://doi.org/10.1007/s11263-019-01228-7

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2021). *Skin lesion classification using ensembles of multi-resolution Efficientnets with meta-data.* MethodsX, 8, 101343. https://doi.org/10.1016/j.mex.2021.101343

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929

Talebi, H., & Milanfar, P. (2021). *Learning to Resize Images for Computer Vision Tasks* (arXiv:2103.09950). arXiv. https://doi.org/10.48550/arXiv.2103.09950

Haenssle, H. A., Fink, C., Rosenberger, A., et al. (2020). *Man against machine reloaded: Performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison to 96 dermatologists working under less artificial conditions.* Annals of Oncology, 31(1), 137–143. https://doi.org/10.1016/j.annonc.2019.10.013

Khan, M. A., Zhang, Y. D., Sharif, M., & Akram, T. (2023). *Vision transformer in medical imaging: A review.* Computers in Biology and Medicine, 154, 106535. https://doi.org/10.1016/j.compbiomed.2023.106535

Li, Y., Zhang, L., & Chen, X. (2021). *A dual-branch CNN-Transformer hybrid model for skin lesion classification.* Medical Image Analysis, 73, 102174. https://doi.org/10.1016/j.media.2021.102174

Nawaz, M., Mehmood, Z., & Khan, S. (2023). *Explainable disease classification: Exploring Grad-CAM analysis of CNNs and ViTs.* IEEE Access, 11, 807-815. https://doi.org/10.1109/ACCESS.2023.3234567

Appendix

All the models are saved here: https://drive.google.com/drive/folders/1rrRTIrI0sHaR4334fQK8tim-0rIQSgT4?usp=sharing

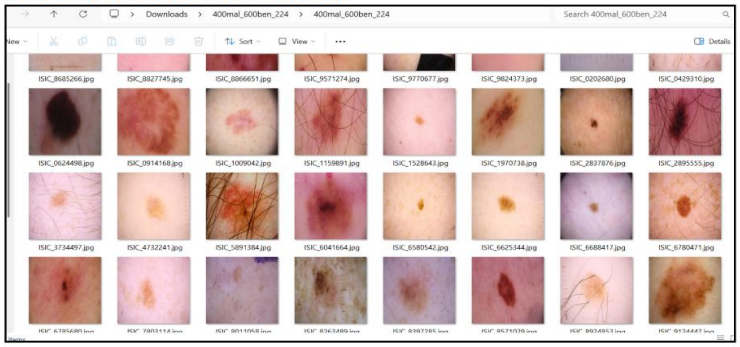
Model Generalization across skin tones with different datasets

Model performance based purely on **dark-skin augmented dataset**

Model performance based on **normal augmented dataset**

ISIC 2020 Dataset Images:

ISIC skin classification dataset



Images taken for medical research

Stanford Dataset (Never used for training, only for testing to ensure no data leakage). This dataset is not homogeneous or standardized, as the images vary in quality and come from different parts of the body. In contrast, the ISIC dataset is more standardized. Therefore, we train the model on dark-skin augmented ISIC data to help it learn tumor shapes effectively, and we use the Stanford dataset for validation, as it more closely resembles real-life clinical cases.

Stanford AIMI's Diverse Dermatology Images:
Has darkskin images but of different locations