# Skin Lesion Classification

AIT Deep Learning Spring 2025: Medical AI Sorcerers

Jonah Levine
*AIT Deep Learning Spring 2025*
*Haverford College*
jblevine@haverford.edu

Ethan Sandoval
*AIT Deep Learning Spring 2025*
*Harvey Mudd College*
etsandoval@g.hmc.edu

Kevin Chen
*AIT Deep Learning Spring 2025*
*Harvey Mudd College*
kevinchen@g.hmc.edu

*Abstract*—**This project investigates the application of deep learning for binary and multiclass classification of skin lesions to aid in early detection of skin cancer. Using high-quality dermatoscopic images and metadata from the ISIC dataset, we trained convolutional neural network models, including a MobileNetV2-based classifier fine-tuned with global average pooling and binary cross-entropy loss. Through balanced training data and model evaluation, the approach attempts to demonstrate the potential to improve melanoma detection by using both image data and structured metadata.**

## I. Introduction

Skin cancer is one of the most common forms of cancer worldwide and its early detection is critical to improving patient outcomes. Traditional diagnostic methods often rely on visual examination followed by biopsy, which can be resource intensive and depend on the availability of experts. To address these limitations, recent research has increasingly explored the use of deep learning to automate **skin lesion classification**, with the aim of enabling faster, more accurate, and accessible diagnosis.

Previous studies have demonstrated the effectiveness of **convolutional neural networks (CNNs)** in classifying skin lesions, utilizing both dermatoscopic images and patient metadata. For example, hybrid architectures such as **U-Net with MobileNetV3** have achieved state-of-the-art accuracy, while other models have explored custom CNNs with data augmentation and transfer learning strategies. However, these models often suffer from issues such as limited generalizability and high computational cost.

This project builds upon these solutions by integrating a light-weight, image-based deep learning model with structured metadata inputs, aiming to capture both accuracy and efficiency. Using the **ISIC dataset**, we investigate a **MobileNetV2**-based architecture fine-tuned for binary and multiclass classification of skin lesions. By addressing common challenges such as missing metadata and overfitting, we aim to contribute a reliable and reproducible tool to enhance melanoma detection and support clinical decision-making.

## II. Previous Work

Recent advancements in skin cancer detection using deep learning have demonstrated promising results, with a variety of architectural and optimization strategies.

### A. Melanoma Detection

One such project by Github user *kshitij-raj* features a custom CNN with data augmentation to address class imbalance [1]. The model featured common deep learning components (e.g., convolutional, pooling, dropout, and dense layers) and emphasized quick, early diagnosis to potentially replace time-consuming biopsy workflows.

### B. U-Net and MobileNetV3

Another implementation by *Lilhore et al.* (2024) proposed a hybrid U-Net and Improved MobileNet-V3 model, optimized using Bayesian search [2]. Their model incorporated dilated convolutions and bias loss functions to outperform standard architectures, achieving 98.86 percent accuracy on the HAM10000 dataset.

### C. Deepskin

Additionally, the DeepSkin model explored transfer learning with pre-trained CNNs like VGG16 and ResNet50, achieving significant performance gains [3]. It highlighted the importance of data preprocessing and layer-wise training for effective classification, noting the role of segmentation in boosting overall accuracy.

Despite their contributions, these approaches often rely on highly complex pipelines or expensive computational resources. Our project aims to build upon their successes by focusing on simplicity and reproducibility, while also exploring the limitations of current solutions in terms of model interpretability.

## III. Dataset

For our project, we used the **International Skin Imaging Collaboration (ISIC) Archive**, a publicly available dataset containing over 500,000 high-quality dermoscopic images of skin lesions. ISIC is developing proposed digital imaging standards and engages the dermatology and computer science communities to improve diagnostic accuracy with the aid of AI.

We selected this dataset for several key reasons:

- **Expert-verified labels**: The images are annotated and reviewed by dermatology professionals, providing a high level of reliability and clinical relevance. Furthermore,

each image is accompanied by a ground truth label indicating whether the lesion is *benign* or *malignant*.
- **High-quality medical images**: The images in the ISIC Archive are high-resolution dermoscopic photographs of skin lesions, taken using standardized imaging techniques commonly used in dermatology clinics.
- **Publicly accessible API**: ISIC offers an open API for downloading and exploring the dataset, making it easy to integrate into research workflows and reproduce experiments.

It is important to note that the ISIC dataset is highly imbalanced, with approximately **96%** of the images labeled as benign. This means that malignant lesions are significantly underrepresented. While this class imbalance poses a challenge for training our deep learning model, it also reflects the real-world distribution of skin lesions encountered in clinical settings. To address the effects of this imbalance and improve model generalization, we experimented with data augmentation strategies and constructed both balanced and unbalanced training subsets.

### A. Preprocessing

*Images*

To prepare the images for training, we performed the following preprocessing steps:

- **Image resizing**: All images were resized to 128 × 128 pixels to reduce computational requirements while preserving key visual features.
- **Data augmentation (for specific experiments)**: For some model variants, we applied random augmentations including flipping, rotation, and zooming to increase training set variability and reduce overfitting.

*Metadata*

The metadata provided with image included columns for 36 variables. Some of these metadata columns included: `age_approx`, `anatom_site_general`, `fitzpatrick_skin_type`, `clin_size_long_diam_mm`, along with general info such as pixel dimensions or patient ID numbers. Unfortunately, most of these columns were filled with NaN values.

We ended up filtering the metadata down to three primary variables for training: **age_approx**, **sex**, **clin_size_long_diam_mm** (longest diameter length of lesion). Additionally, we also used the **diagnosis_1** column for our target variable, as this indicated whether the associated lesion was *benign* or *malignant*.

| **Feature** | **Description** |
|---|---|
| `age_approx` | Approximate age of the patient. |
| `sex` | Sex of the patient. |
| `clin_size_long_diam_mm` | Longest diameter length of the lesion in millimeters. |
| `diagnosis_1` | Indicates whether the lesion is *benign* or *malignant*. (Target Variable) |

TABLE I: Description of selected dataset features.

To address missing values within these three main metadata features we dropped rows without `sex` information, and filled in missing `clin_size_long_diam_mm` and `age_approx` cells with their respective mean values.

## IV. Method

We focused on a **binary classification task**: predicting whether a lesion is *benign* or *malignant*. To better control for factors like class imbalance and model robustness, we created multiple dataset variations, including balanced subsets, real-world class distributions, and augmented samples. As mentioned above, all images were resized to 128 × 128 pixels to optimize training speed without significantly sacrificing detail.

### A. Baseline Models

Before employing more advanced architectures, we established baseline models using both image and metadata features to evaluate the relative effectiveness of different input types.

*1) Metadata Linear Regression:* As a simple baseline, we trained a logistic regression model using only tabular metadata (e.g., age, sex, anatomical site). This model served as a benchmark to evaluate how much predictive power exists in the non-image features alone.

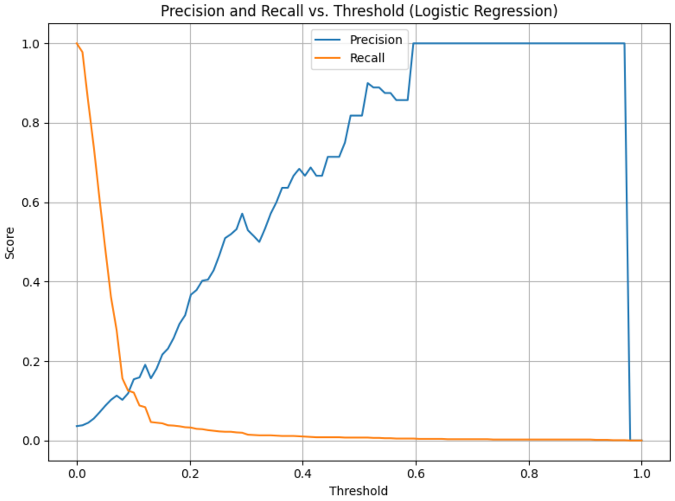
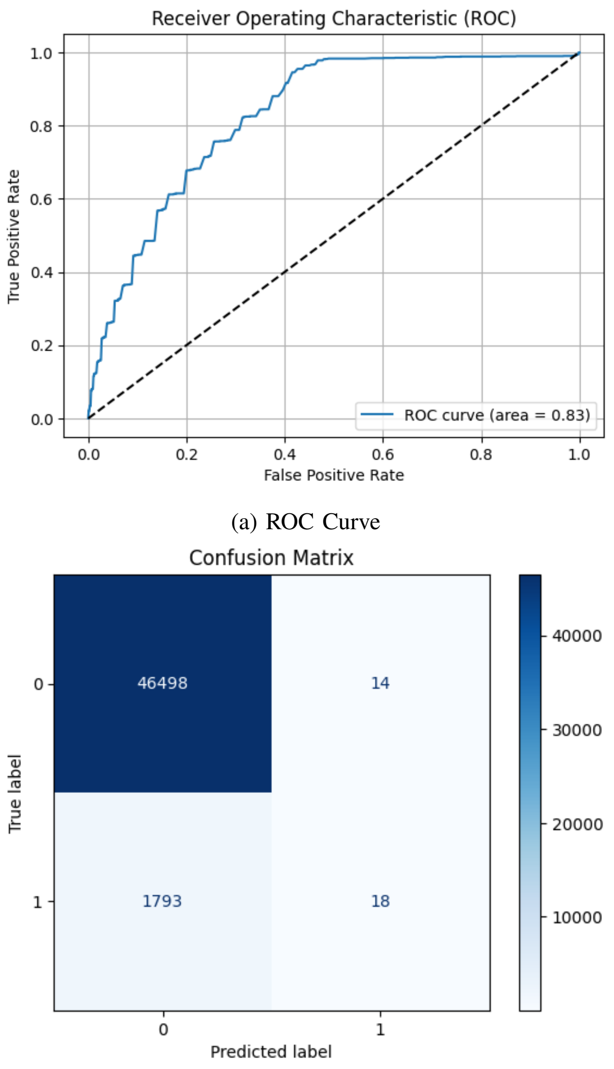


Fig. 1: Precision-Recall Curve for Linear Regression on Image Metadata.

*2) Metadata MLP:* To better capture potential nonlinear relationships in the metadata, we implemented a multilayer perceptron (MLP). This model consisted of two dense layers with ReLU activations and dropout for regularization.

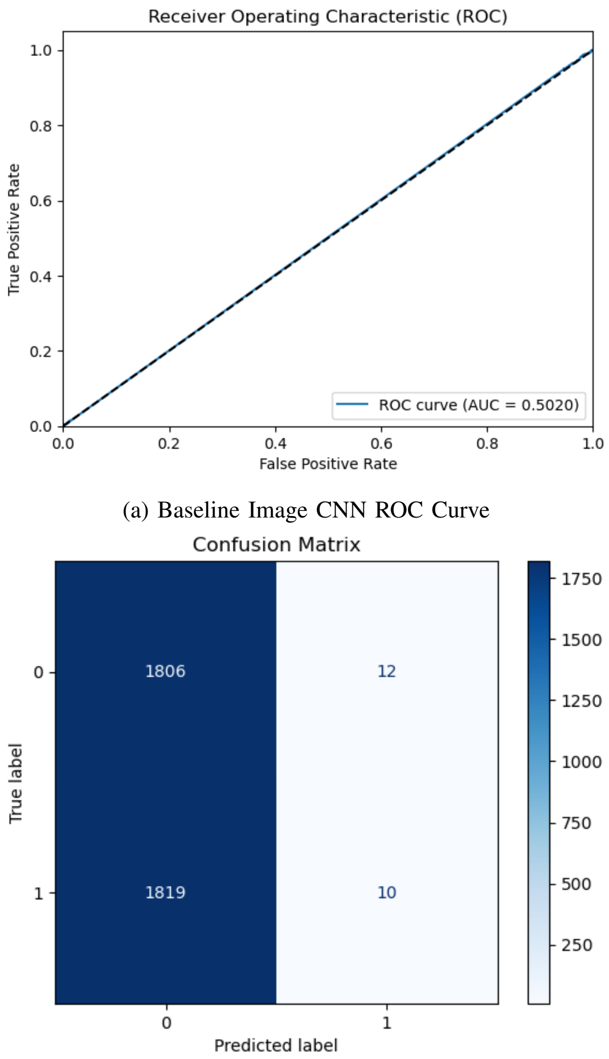
The ROC curve in Figure 2a initially suggests strong performance for the MLP model. However, this is misleading due to the highly imbalanced nature of the test set, which mirrors the original ISIC distribution of approximately 96% benign lesions. As shown in Figure 2b, the model predicted nearly all inputs as *benign*, effectively functioning as a majority-class classifier. While this results in a seemingly high AUC, the



(a) ROC Curve

(b) Confusion Matrix

Fig. 2: Model evaluation using ROC curve and confusion matrix.



(a) Baseline Image CNN ROC Curve

(b) Baseline Image CNN Confusion Matrix

Fig. 3: Baseline Image CNN model evaluation using ROC curve and confusion matrix.

model has a minimal practical utility for detecting malignant cases.

*3) Image CNN:* We also implemented a basic convolutional neural network (CNN) trained directly on image inputs. This network included several convolutional layers with max pooling, followed by flattening into dense layers.

While this model was able to learn from a more intricate dataset consisting of images, when tested on a balanced test set, it was essentially as effective as a random classifier, as can be seen in Figure 3.

### *B. Architecture*

Since image classification is an extensively studied problem, we leveraged transfer learning by incorporating a pre-trained base model and customizing the upper layers for our binary classification task.

*1) Model Selection: MobileNetV2:* For our deep learning architecture, we selected **MobileNetV2** as the base model for skin lesion classification due to its efficiency, performance, and suitability for transfer learning. Below are the key reasons motivating this choice:

- **Lightweight Architecture**: MobileNetV2 is designed for resource-constrained environments, making it highly efficient in terms of memory and computational requirements. This enabled faster training and experimentation since we were limited to our personal hardware.
- **Depthwise Separable Convolutions**: A core feature of MobileNetV2 is the use of depthwise separable convolutions, which significantly reduce the number of parameters and operations while maintaining high representational power.
- **Transfer Learning Capabilities**: We utilized a Mo-

bileNetV2 model pre-trained on the ImageNet dataset and fine-tuned it for our binary classification task. This approach allowed us to leverage rich feature representations learned from large-scale natural images, which transfer effectively to medical imaging tasks.
- **Strong Performance on Small Datasets**: Given that medical image datasets are often limited in size, MobileNetV2's ability to perform well in low-data regimes made it a suitable candidate for our project.
- **Established Use in Medical Applications**: MobileNetV2 has been successfully applied in various biomedical imaging problems, demonstrating its adaptability and effectiveness in clinical contexts such as skin cancer detection.

*2) Additional Architecture:* To adapt MobileNetV2 to our specific task, we appended a custom classification head to the base model. This included a **2D Global Average Pooling (GAP)** layer, followed by dense layers for classification.

We selected GAP instead of flattening or fully connected layers immediately after convolutional layers for several reasons:

- **Parameter Efficiency**: GAP dramatically reduces the number of trainable parameters by replacing high-dimensional feature maps with a single value per channel. This reduces the risk of overfitting—especially important for medical datasets, which often have limited sample sizes.
- **Spatial Invariance**: By averaging across the spatial dimensions of feature maps, GAP encourages the model to focus on *what* is present rather than *where*—a useful property in lesion detection, where the presence of features such as irregular borders or color variation is more important than their exact location.
- **Interpretability and Robustness**: GAP is often associated with better interpretability and smoother training. It acts as a structural regularizer by forcing the network to extract global features that generalize better, which is crucial for clinical tasks requiring reliable and robust predictions.
- **Proven Performance**: GAP has been successfully used in many medical imaging studies and is a common design choice in modern convolutional neural networks such as ResNet and MobileNet variants.

Following the GAP layer, we added a fully connected dense layer with ReLU activation and a final sigmoid-activated output layer for binary classification.

In some model variants, we also incorporated **metadata features** (e.g., patient age, sex, anatomical site). These were concatenated with the image-based feature vector after the pooling operation, allowing the network to integrate both visual and contextual clinical information.

### *C. Training*

The model was trained using the following hyperparameters and techniques:

We trained separate model variants using different data balancing strategies (balanced, unbalanced, augmented), allowing us to evaluate the impact of data distribution and augmentation on performance.

| **Parameter** | **Value** |
|---|---|
| Learning Rate | 0.001 |
| Optimizer | Adam |
| Loss Function | Binary Cross-Entropy |
| Batch Size | 32 |
| Epochs | 50 (with Early Stopping) |
| Early Stopping Patience | 10 |
| Image Size | $128 \times 128$ |
| Regularization | Early stopping based on validation loss |

TABLE II: Training hyperparameters and configuration.

## V. Evaluation

We conducted a series of experiments to evaluate the performance of our skin lesion classification model, beginning with **binary classification** and progressing to **multiclass classification** with increasing model complexity and data augmentation.

First, we trained a binary classification model using a **balanced dataset** containing 5000 *benign* and 5000 *malignant* cases. This allowed for fair evaluation across both classes. We assessed the model using a **confusion matrix** and a **Receiver Operating Characteristic (ROC) curve**, providing insight into both *class-specific performance* and overall *discriminative ability*. We again performed the same test using subsets of 10000 images each.

Next, we maintained a balanced training set but evaluated the model using the **original unbalanced validation set** (approximately **9400 benign** images and **600 malignant** images). This setting more accurately reflected *real-world distributions*. Again, we generated a confusion matrix and ROC curve to assess model performance under **class imbalance**.

To improve *generalization* and combat *overfitting*, we repeated the above setup using **augmented image data** during training. This included transformations such as *flipping*, *rotation*, and *zooming*. Using the augmented data, we experimented with keeping the convolutional base frozen and unfrozen. The resulting models were evaluated using both a confusion matrix and an ROC curve.

We then implemented a **pre-trained ResNet50** architecture in place of MobileNetV2, fine-tuning the top layers for binary classification. This configuration was evaluated using a confusion matrix.

Building on this, we incorporated **metadata features** (e.g., *age*, *patient sex*, *lesion size*) alongside image data using a **dual-input model**. We evaluated this model using a confusion matrix and ROC curve.

Finally, we extended the task to **multiclass classification**, differentiating between different types of *malignant* lesions, including *basel cell carcinoma*, *melanoma*, *melanoma metastasis*, and *squamous cell carcinoma*. We trained two variants: one with the **pre-trained base model frozen** and another with the base model **unfrozen for fine-tuning**. Each configuration was evaluated with a confusion matrix, providing a detailed view of classification performance across all classes.

## VI. Results

Below are some figures that capture the **test results** of our evaluations.

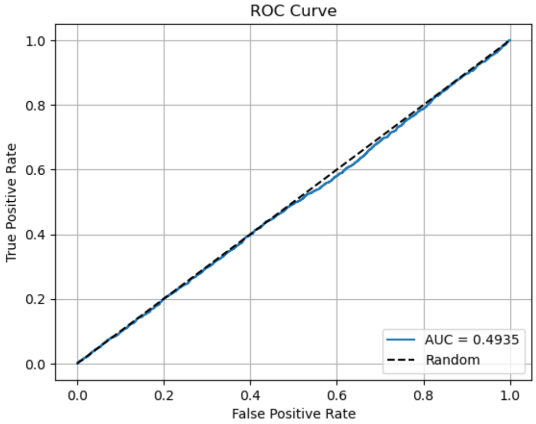


Fig. 4: ROC curve using balanced data for training and evaluation

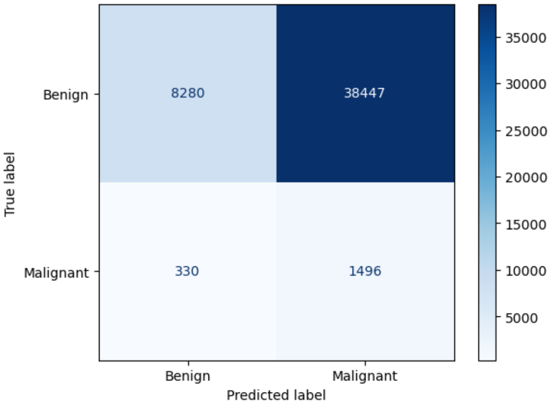


Fig. 5: Confusion matrix using balanced data for training and evaluation

## VII. Discussion

The initial evaluation was conducted using **balanced image data** for both **training and validation**. As illustrated in **Figure 4**, the resulting ROC curve yielded an **AUC of 0.4935**, which is only marginally better than random guessing. This indicates that the model is not effectively distinguishing between *benign* and *malignant* tumors at this stage.

Further evidence of this performance is shown in the confusion matrix (**Figure 5**). The model predicted **38,447 benign cases** as malignant and **330 malignant cases** as benign. Despite the data being perfectly balanced, the model demonstrates a significant bias toward the majority class in its predictions, suggesting that overfitting had occured. This is despite the model achieving relatively solid results during validation, achieving 84.55 percent accuracy (**Figure 6**). Nearly

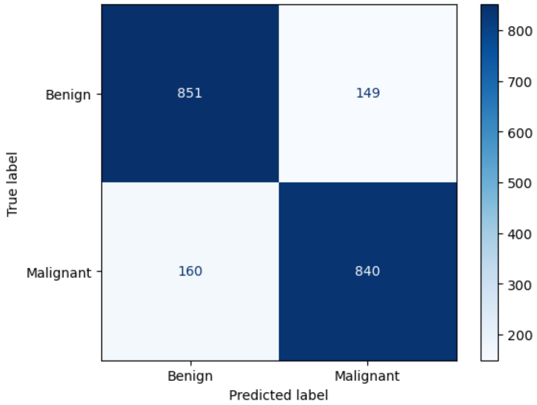


Fig. 6: Confusion matrix for validation set in balanced data experiment

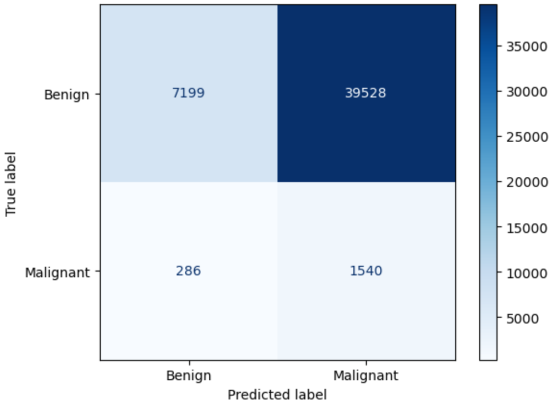


Fig. 7: Confusion matrix using balanced data with a subset of 10000 images

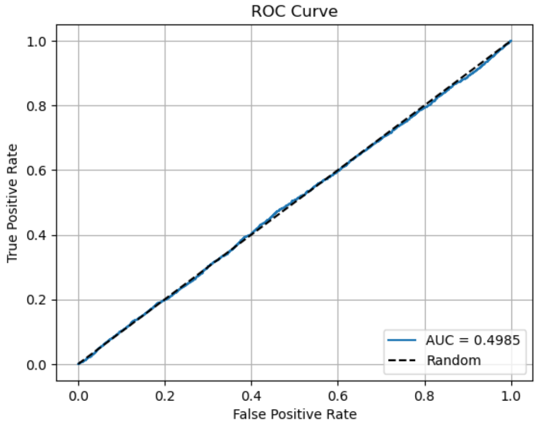


Fig. 8: ROC curve using original data splits for evaluation

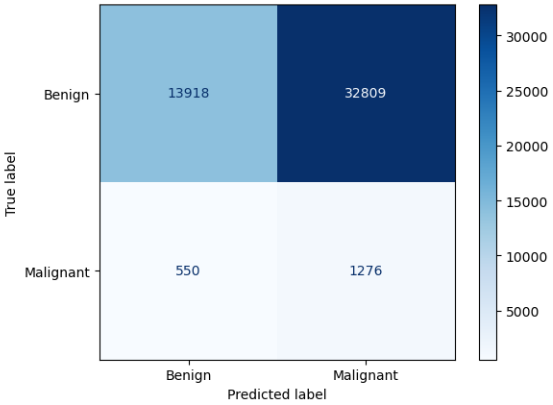


Fig. 9: Confusion matrix using original data splits for evaluation

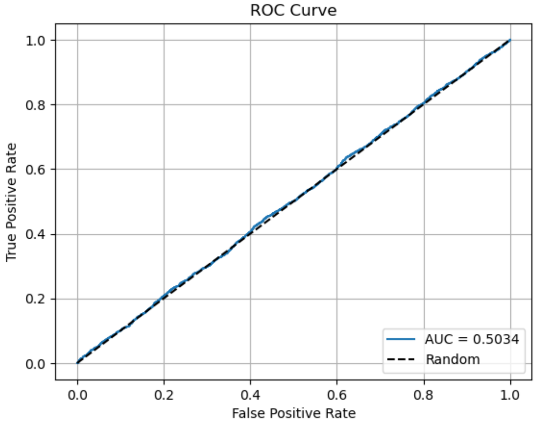


Fig. 10: ROC curve using augmented and balanced image data for training and unbalanced data splits data splits for evaluation

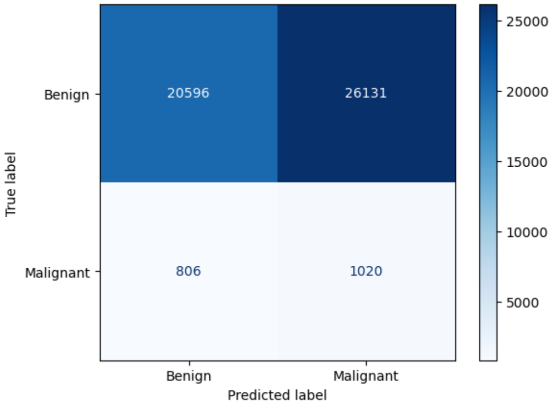


Fig. 11: Confusion matrix using augmented and balanced image data for training and unbalanced data splits data splits for evaluation

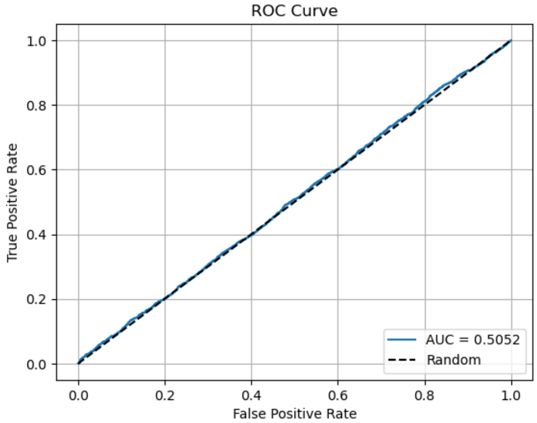


Fig. 12: ROC curve using augmented image data and freezing

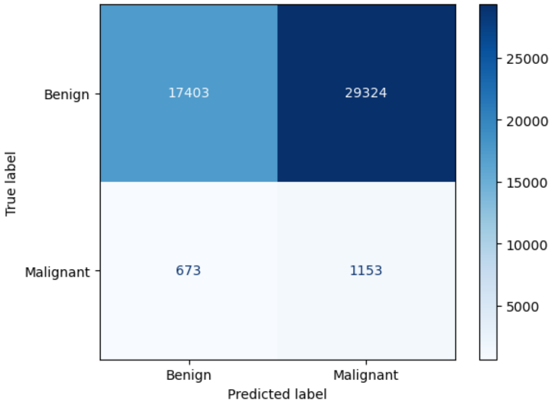


Fig. 13: Confusion matrix using augmented image data and freezing

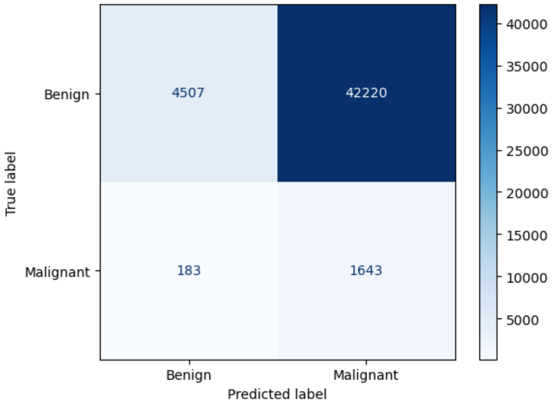


Fig. 14: Confusion matrix evaluated using RESNET50 pre-trained base model instead of MobilenetV2

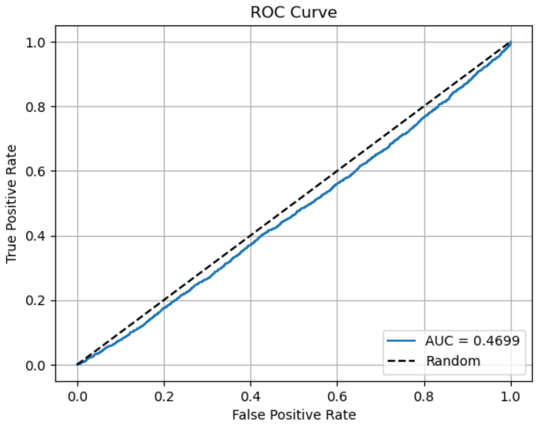


Fig. 15: ROC curve with metadata added

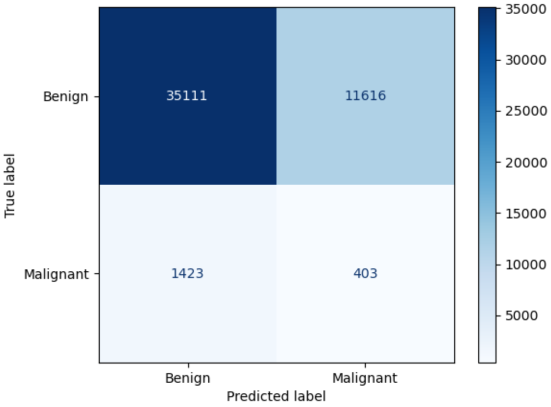


Fig. 16: Confusion matrix with metadata added

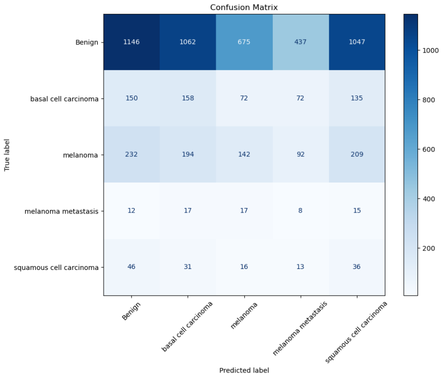


Fig. 17: Multiclass confusion matrix with no unfreezing

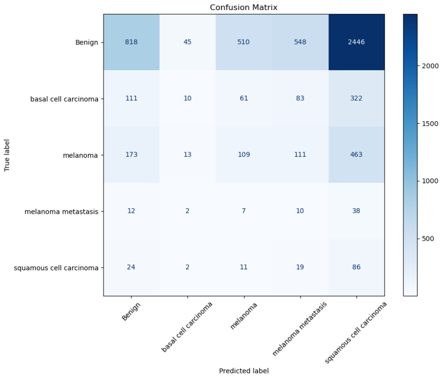


Fig. 18: Multiclass confusion matrix with unfreezing

identical results can be seen in the same test conducted with a balanced training subset of 10,000 images (**Figure 7**).

The second evaluation experiment involved training on **balanced image data**, but evaluating on the **unbalanced original distribution** of the dataset, where *benign cases comprise approximately 96%* of the validation set. The ROC curve in **Figure 8** shows an **AUC of 0.4985**, again akin to random guessing.

The corresponding confusion matrix (**Figure 9**) further illustrates the model's difficulty in generalizing to the skewed class distribution. While the number of true positives for the malignant class remains comparable (1276), there is a notable increase in false negatives (550 malignant cases misclassified as benign), and a very high false positive count (32,809 benign cases misclassified as malignant). This outcome suggests that although the model learned from a balanced distribution, it failed to adapt to the natural skew of real-world clinical data.

To further address potential overfitting and limited generalizability, we applied a series of ***image augmentations*** during training, including random flipping, rotation, and zoom. Using this ***augmented and balanced training set***, we evaluated the model on the original, ***unbalanced validation data***.

As shown in **Figure 10**, the ROC curve yielded an **AUC of 0.5034**, indicating slightly improved but still near-random classification performance. The corresponding confusion matrix in **Figure 11** highlights ongoing difficulty in differentiating malignant samples, though a slight increase in true positives (malignant detected as malignant) is observed compared to the previous unaugmented training run.

### *Effect of Data Augmentation Without Unfreezing the Base Model*

An additional experiment explored the impact of data augmentation alone by training a MobileNetV2-based classifier with the convolutional base frozen. As seen in **Figure 12**, despite employing horizontal flips, small rotations, and zooming, the model still achieved an AUC of only **0.5052**, yet again barely above random chance. The confusion matrix (**Figure 13**) also reveals significant class confusion, particularly with a high false positive rate for malignant cases.

These results indicate that augmentation may provide some benefit in learning more generalizable features, but are still not sufficient to overcome the limitations of the base model or the challenges posed by extreme class imbalance in the evaluation data.

In an attempt to improve performance beyond the capabilities of MobileNetV2, we employed a ***ResNet50-based architecture*** with ***pretrained ImageNet weights*** as the feature extraction backbone, trained on the balanced data but evaluated on the unbalanced data.

As shown in **Figure 14**, the model demonstrated a **notable improvement** in correctly identifying malignant cases, with 1,643 true positives compared to 1,020 in the previous augmented CNN. Moreover, the number of false negatives (malignant predicted as benign) was reduced to 183, indicating a stronger capacity to distinguish high-risk cases. However, it only achieved this because it identified nearly everything as malignant, including most of the benign images.

Next, we trained the model using **balanced image data** and accompanying *metadata* (such as age and anatomical site), then evaluated on an **unbalanced validation set**. The ROC curve in **Figure 15** shows an AUC of **0.4699**, which is *below random performance*, indicating that the model struggled to extract useful discriminative features even with metadata incorporated.

The confusion matrix in **Figure 16** supports this conclusion: while benign samples were mostly classified correctly, malignant examples were misclassified at a high rate, with **1,423** false negatives and only **403** true positives. This result suggests that the model failed to effectively concatenate the additional metadata under the given architecture and training setup.

To extend our analysis beyond binary classification, we applied our models to a multiclass setting, differentiating between *benign*, *basal cell carcinoma*, *melanoma*, *melanoma metastasis*, and *squamous cell carcinoma*. We trained two variants: one with the **pre-trained base model frozen**, and one where the base model was **unfrozen for fine-tuning**.

**Figure 17** presents the confusion matrix for the frozen model. The model exhibited a strong bias toward predicting the majority class, *benign*, with substantial misclassification across all malignant categories. In particular, the model incorrectly classified a large portion of malignant lesions (e.g., melanoma and squamous cell carcinoma) as benign or squamous cell carcinoma, perhaps a sign of poor separation of features.

In contrast, **Figure 18** shows the confusion matrix for the *unfrozen* model. Here, we observe notable improvements in capturing more meaningful distinctions across the malignant classes. For instance, predictions of melanoma and basal cell carcinoma were more evenly distributed, and benign lesions were less frequently confused with malignancies.

These results highlight the benefit of *fine-tuning* the base model in a multiclass clinical setting, as it enables the model to adapt better to subtle inter-class visual features, improving **granular classification performance**.

## VIII. Summary

This project explored the use of deep learning for automated classification of skin lesions, with the goal of improving the early detection of skin cancer. We trained a series of binary and multiclass models on dermoscopic images from the ISIC dataset, experimenting with different data balancing strategies, image augmentations, architectural choices, and the integration of metadata.

Initial experiments using a basic CNN and balanced datasets resulted in near-random classification performance, with AUC scores below 0.5 and confusion matrices indicating poor generalization. Using unbalanced evaluation sets further exposed model biases and high false positive rates. Data augmentation slightly improved sensitivity to malignant cases, but was insufficient to produce high quality performance.

Little improvements were achieved through the use of a pretrained ResNet50 model. Although it saw substantially reduced false negatives and increased correct malignant classification, it identified the vast majority of benign images as malignant. Incorporating metadata under a basic concatenation architecture failed to yield gains, suggesting the need for more advanced feature fusion techniques. Multiclass classification further benefited from fine-tuning the base model, demonstrating improved differentiation among lesion types when the pretrained layers were unfrozen.

Overall, our findings suggest that while deep learning holds promise for skin lesion classification, careful attention must be paid to model architecture, class balance, and data representation. Challenges remain in maximizing diagnostic accuracy for underrepresented classes and leveraging multimodal inputs.

## References

[1] K. Raj, “Melanoma Skin Cancer Detection: Build a CNN-based model which can accurately detect melanoma,” *GitHub*, 2021. [Online]. Available: https://github.com/kshitij-raj/Melanoma-Skin-Cancer-Detection

[2] U. K. Lilhore, S. Simaiya, Y. K. Sharma *et al.*, “A precise model for skin cancer diagnosis using hybrid U-Net and improved MobileNet-V3 with hyperparameters optimization,” *Scientific Reports*, vol. 14, no. 1, p. 4299, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-54212-8

[3] H. L. Gururaj, N. Manju, A. Nagarjun, V. N. M. Aradhya and F. Flammini, “DeepSkin: A Deep Learning Approach for Skin Cancer Classification,” *IEEE Access*, vol. 11, pp. 50205–50214, 2023. doi: 10.1109/ACCESS.2023.3274848

[4] Keras Team, “Keras Applications: ResNet and ResNetV2,” 2024. [Online]. Available: https://keras.io/api/applications/resnet/

[5] Keras Team, “Keras Applications: MobileNet, MobileNetV2, and MobileNetV3,” 2024. [Online]. Available: https://keras.io/api/applications/mobilenet/