SKIN CANCER CLASSIFICATION USING DEEP LEARNING MODELS PROGRESS REPORT

Vishwas Puri
Student# 1010174669
vishwas.puri@mail.utoronto.ca

Rayan Ahsan
Student# 1010140820
rayan.ahsan@mail.utoronto.ca

Adyan Hossain
Student# 1010150373
adyan.hossain@mail.utoronto.ca

Kashan Ahmad
Student# 1010090072
kashan.ahmad@mail.utoronto.ca

ABSTRACT

This document presents the progress report for Group 29's APS360 final project on classifying various types of skin lesions using deep learning. The report outlines the current status of our baseline and primary models, including implementation details, performance results, and encountered challenges. We describe the use of convolutional neural networks (CNNs) and a pretrained ResNet18 model to improve performance and training efficiency. Quantitative and qualitative results are provided to demonstrate the feasibility of our approach. The report also includes updated project goals, individual team member contributions, and a plan to complete the remaining milestones. —-Total Pages: 9

1 BRIEF PROJECT DESCRIPTION

Skin cancer is among the most common forms of cancer worldwide, with approximately 80,000 cases diagnosed annually in Canada Canadian Skin Cancer Foundation (2025). Early and accurate recognition of skin cancer significantly improves outcomes, with a five-year survival rate of 99% Skin Cancer Foundation (2025). However, dermatologist availability, especially in rural areas, remains limited.

To address this gap, this project develops a deep learning model to classify dermoscopic images of skin lesions into seven diagnostic categories using the HAM10000 dataset. Convolutional Neural Networks (CNNs), a type of deep learning model designed to process visual data by automatically learning hierarchical features, are well-suited for this task. They can capture discriminative visual patterns such as colour variation and border irregularity directly from raw images, enabling improved diagnostic accuracy and consistency.

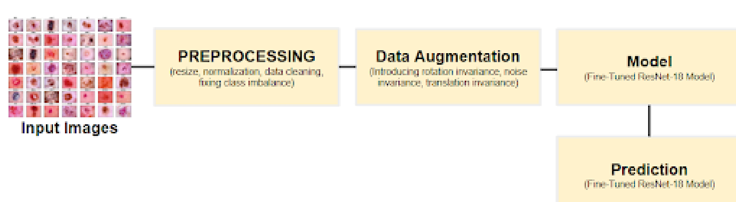

Figure 1: Flowchart displaying projects pipeline

The classification pipeline takes a dermoscopic image as input, applies preprocessing and data augmentation, then feeds the image into a CNN model (ResNet-18). The model outputs a predicted lesion category.

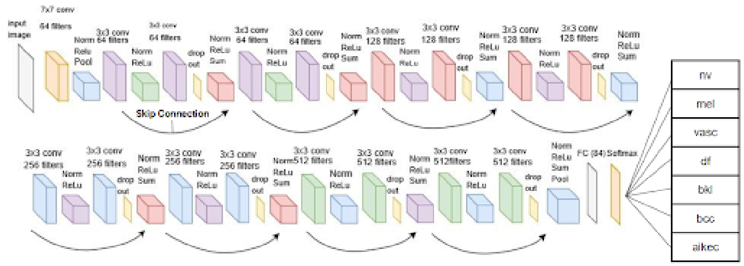

Figure 2: ResNet 18 architecture

The model used is ResNet-18, a convolutional neural network that processes images through an initial convolutional layer and a series of residual blocks with skip connections.

2 Individual Contributions and Responsibilities

Our team operates with a strong emphasis on collaborative development and clear division of responsibilities to ensure efficient progress towards our project goals. We leverage GitHub Projects for task assignment, progress tracking, and milestone management. Discord facilitates real-time communication and our weekly meetings. Our primary platform for collaborative code development and sharing is Google Colab, while GitHub is used for version control. Collaborative writing and documentation are handled through Google Docs.

2.1 Team Member Roles and Accomplishments

Vishwas Puri
Responsibilities: Data Acquisition & Preprocessing, Primary Model Training & Statistics, Feature Analysis, Application Deployment.
Accomplishments: Led data acquisition and preprocessing efforts for the HAM10000 dataset, and contributed to the training and performance analysis of the primary model.

Adyan Hossain
Responsibilities: Primary Model Adaptation, Model Architecture Refinement, Performance Statistics & Analysis.
Accomplishments: Modified the pretrained ResNet18 model's fully connected layer to match our seven classification classes.

Rayan Ahsan
Responsibilities: Primary Model Implementation Lead (ResNet18), Transfer Learning, Model Optimization, Data Preprocessing & Augmentation, Baseline Model Statistics, Hyperparameter Tuning.
Accomplishments: Contributed to data preprocessing and data augmentation efforts, implemented ResNet18 with transfer learning, conducted initial primary model training runs showing improvements, and performed initial Grad-CAM heatmaps analysis.

Kashan Ahmad
Responsibilities: Baseline Model Development, Baseline Model Statistics, Assist Primary Model Training, Grad-CAM Generation, GPU Resource Investigation.
Accomplishments: Developed and conducted initial training and evaluation of the baseline CNN model and generated quantitative and qualitative results for the baseline CNN model.

2.2 Upcoming Tasks

Date Range	Task	Responsible
July 14th - 21st	Lead tuning of hyperparameters (LR, decay rates, batch size, etc.) on primary models using techniques like random sampling or Naive Bayes, and select the best model from these configurations.	Rayan Ahsan
July 22nd - 25th	Generate detailed performance statistics such as validation/training loss and accuracy curves, recall, precision, f1 score, and confusion matrix, comparing them to our model before hyperparameter tuning.	Adyan Hossain
July 25th - 27th	Employ Gradient-weighted Class Activation Mapping to generate heatmaps of the images indicating how much certain geometric features contribute to the classification decision..	Kashan Ahmad
July 27th - 30th	Compare Grad-CAM heatmaps with actual features dermatologists use to classify skin lesions (i.e., border irregularities, color, etc.) and make modifications if necessary.	Vishwas Puri
July 30th - August 5th	Deploy the application as a website using Streamlit to allow user image uploads and lesion classification.	Vishwas Puri
July 30th - August 5th	Investigate external GPU resources for extended model training.	Kashan Ahmad
August 6th - 19th	Collaborate on final Project Report and Presentation.	All Members

Table 1: Summary of upcoming tasks and responsible members

2.3 Risk Mitigation and Redundancy

We mitigate risks through cross-training for key responsibilities, maintaining a shared knowledge base (GitHub, Google Docs), conducting regular code reviews, and assigning backup roles for sensitive tasks. Flexible scheduling and asynchronous tools address busy schedules. Conflicts are resolved by documented decisions and team votes. Performance challenges (training time, low accuracy) will be addressed via hyperparameter tuning and exploring additional computational resources (e.g., Google Colab Pro).

2.4 Overall Team Progress Assessment

The team is on track to meet final project objectives. We have completed data acquisition/preprocessing, established a baseline, and implemented a promising primary model. Despite ongoing challenges with computational power and class imbalance, our updated plan provides a clear path to improve accuracy further and deliver a comprehensive solution within the timeframe.

3 Data Processing

This project uses the publicly available HAM10000 dataset, which contains 10,015 dermoscopic images collected from clinical sites in Austria and Australia. The dataset is available on the Harvard Dataverse and is widely used for skin lesion classification tasks.

Each image is labelled with one of seven skin lesion types:

- nv: Melanocytic nevi
- mel: Melanoma
- bkl: Benign keratosis-like lesions

- bcc: Basal cell carcinoma
- akiec: Actinic keratoses and intraepithelial carcinoma
- vasc: Vascular lesions
- df: Dermatofibroma

The distribution of image samples amongst these classes is shown in Figure 3

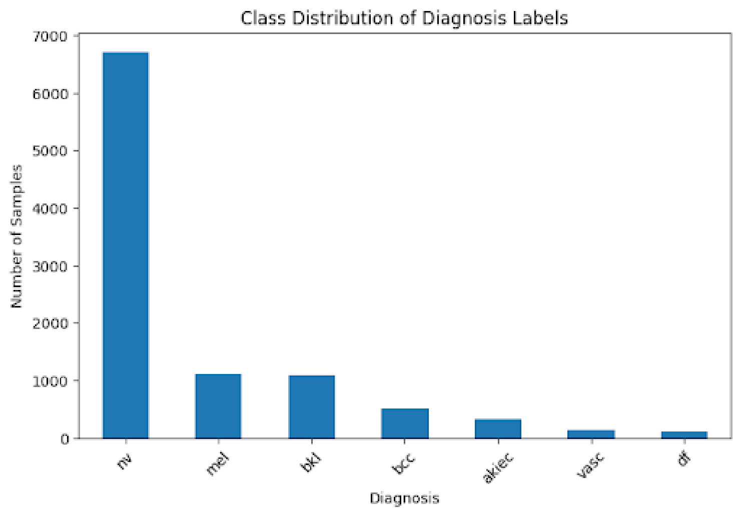

Figure 3: Image displaying the sample distribution amongst classes

The metadata consists of patient information such as age, sex, anatomical site of lesion, and lesion ID.

3.1 Data Cleaning

We cleaned the HAM10000 dataset by removing corrupt, duplicate images and any null values, ensuring only high-quality samples remained. All images were resized to 224×224 and normalized to standardize input for the model. The in-depth process is outlined in Figure 4.

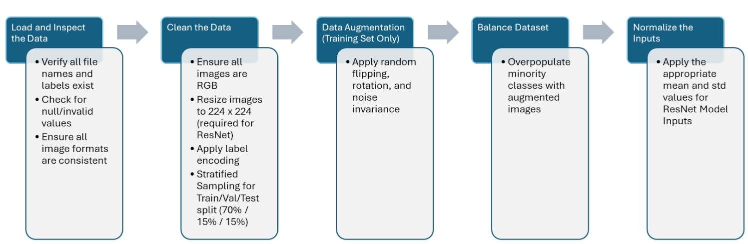

Figure 4: Display of data cleaning and preprocessing process

3.2 Plan for Testing Data

For final evaluation, we will test the model on the held-out test split of HAM10000 (15%). This consists of unseen and unaugmented images to provide the best evaluation of the model's classification ability.

3.3 Challenges

Class Imbalance: As seen in the dataset, the nv class made up over half of the samples, leading to biased predictions toward the majority class. This would have led to poor performance on underrepresented classes like melanoma. We addressed this through oversampling the minority classes via data augmentation.

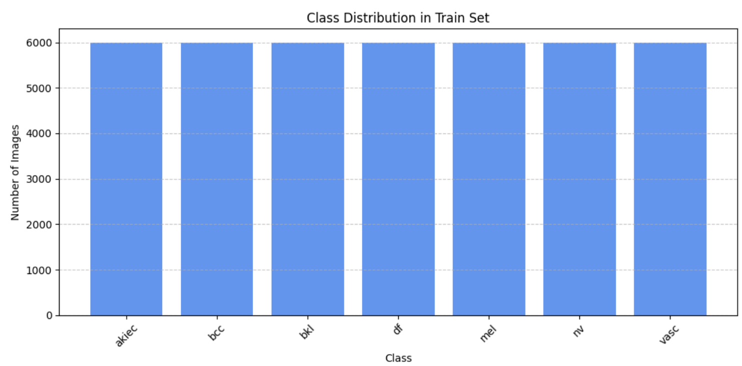

Figure 5: Display of the balanced class distribution following oversampling via data augmentation

Metadata Inconsistency: Metadata fields like age and sex had missing values. To avoid introducing noise or complexity, we excluded those samples from the training pipeline. Our model focuses solely on image data to maintain consistency, reduce potential bias, and avoid inaccurate results.

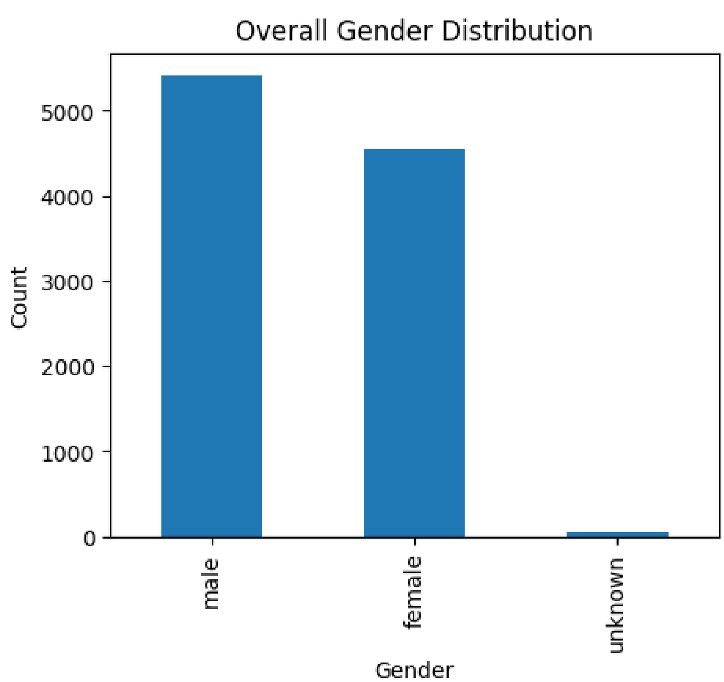

Figure 6: Distribution of dataset samples across genders

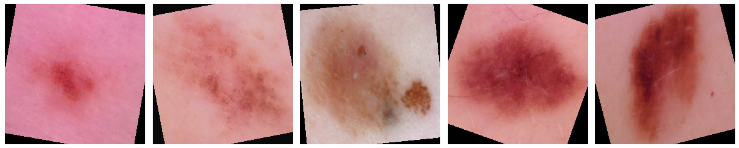

Figure 7: Five image samples following data augmentation

4 BASELINE MODEL

To assess the feasibility of our skin cancer classification project, we implemented a baseline Convolutional Neural Network model. The purpose of this baseline was to provide a simple yet functional deep learning architecture that could be used to benchmark future improvements using more advanced models such as pre-trained networks like ResNet18.

The baseline model was chosen to be a basic encoder-decoder style CNN. The encoder consists of three convolutional layers, with increasing channel depth (32, 64, 128). Each convolutional layer was followed by batch normalization, ReLU activation, and max-pooling operations. The decoder is composed of a flattening layer followed by two fully connected layers, mapping the learned features into a 7-class probability distribution. The architecture was designed to be lightweight and easy to implement, without requiring an extensive amount of hyperparameter tuning.

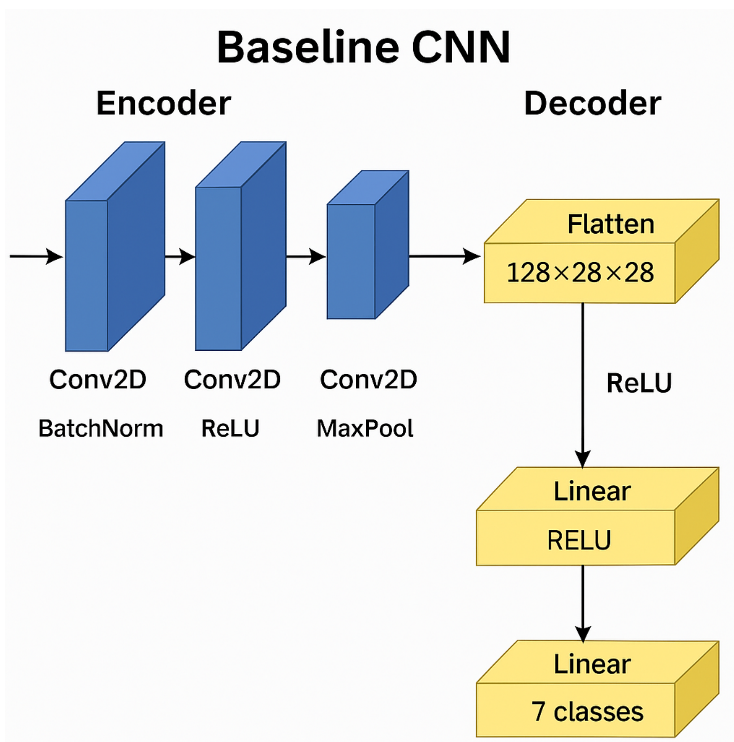

Figure 8: Baseline model architecture

During training, one of the major challenges faced was the size of the dataset. With over 42,000 images (6000 images per class across 7 classes), the training process on both CPU and GPU was extremely slow and not feasible within our limited time constraints. In order to overcome this, we decreased the baseline dataset to approximately 1000 training images and 200 validation images, allowing us to test whether the model could still learn meaningful features from limited samples. Despite the reduced data size, the model was able to achieve a training accuracy of 44.5% and a validation accuracy of 67.5%, as shown in the training log and accuracy plot below.

```
Epoch 1/25 | Loss: 24.6790 | Train Acc: 0.1588 | Val Acc: 0.0150
Epoch 2/25 | Loss: 5.2528 | Train Acc: 0.2387 | Val Acc: 0.5650
Epoch 3/25 | Loss: 2.0317 | Train Acc: 0.2475 | Val Acc: 0.6500
Epoch 4/25 | Loss: 1.7606 | Train Acc: 0.3075 | Val Acc: 0.6600
Epoch 5/25 | Loss: 1.6504 | Train Acc: 0.3275 | Val Acc: 0.6750
Epoch 6/25 | Loss: 1.5919 | Train Acc: 0.3475 | Val Acc: 0.6500
Epoch 7/25 | Loss: 1.4592 | Train Acc: 0.4213 | Val Acc: 0.6600
Epoch 8/25 | Loss: 1.4219 | Train Acc: 0.4450 | Val Acc: 0.6600
Early stopping triggered after 8 epochs.
```

Figure 9: Baseline model training results

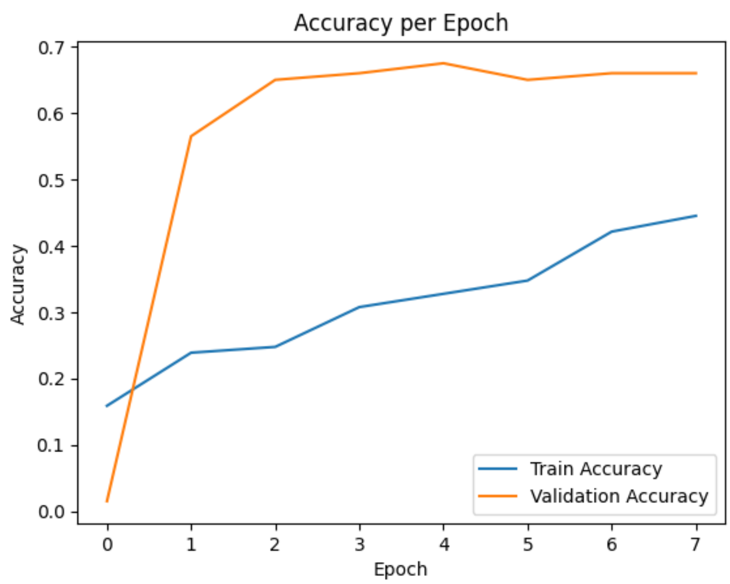

Figure 10: Baseline model validation results

These results suggest that even with limited data, the model was able to identify useful features for skin lesion classifications. The low training accuracy was likely due to the limited number of training samples, restricting the model's ability to fully learn the complex patterns needed. However, the reasonable validation accuracy indicates that the model was not overfitting and had some generalization capability. The model also seemed to accurately predict the nv class however had trouble predicting other classes as seen from the image below. This may be due to the large number of nv classes relative to other classes within the validation dataset.

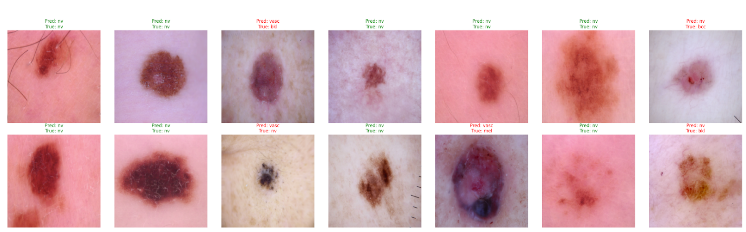

Figure 11: Image showing 14 predicted images with ground truth labels

Overall, this CNN model gives us confidence that if we use a more powerful and efficient architecture, such as a pretrained ResNet18, we can improve performance on the full dataset without compromising feasibility.

5 Primary Model

For our first version of the primary model, we used transfer learning with the pretrained ResNet18 architecture, which is widely recognized for its strong performance on the ImageNet dataset and is a standard backbone for image classification tasks.

ResNet18 consists of 18 layers and introduces skip connections to mitigate accuracy degradation in deeper networks. This architecture allows our model to leverage transfer learning to extract relevant features from our dataset efficiently, making it ideal for initial experimentation. The relatively low complexity also provides flexibility for future fine-tuning as we work towards our final model.

Refer to figure 2 to see how we reconfigured resNet18 to fit our problem where we changed our final fc layer to have the 7 classification output neurons

From this setup we were able to outperform our baseline model by achieving the following training and validation result of 67.12% and 69.71% respectively.

```
Epoch 1/25 | Loss: 1.4433 | Train Acc: 0.4522 | Val Acc: 0.6718
Epoch 2/25 | Loss: 1.1721 | Train Acc: 0.5624 | Val Acc: 0.6738
Epoch 3/25 | Loss: 1.0857 | Train Acc: 0.5920 | Val Acc: 0.6877
Epoch 4/25 | Loss: 1.0097 | Train Acc: 0.6210 | Val Acc: 0.6858
Epoch 5/25 | Loss: 0.9616 | Train Acc: 0.6408 | Val Acc: 0.6971
Epoch 6/25 | Loss: 0.9052 | Train Acc: 0.6614 | Val Acc: 0.6964
Epoch 7/25 | Loss: 0.8759 | Train Acc: 0.6714 | Val Acc: 0.6784
Epoch 8/25 | Loss: 0.8764 | Train Acc: 0.6712 | Val Acc: 0.6798
Early stopping triggered after 8 epochs.
```

Figure 12: Image showing training of primary model

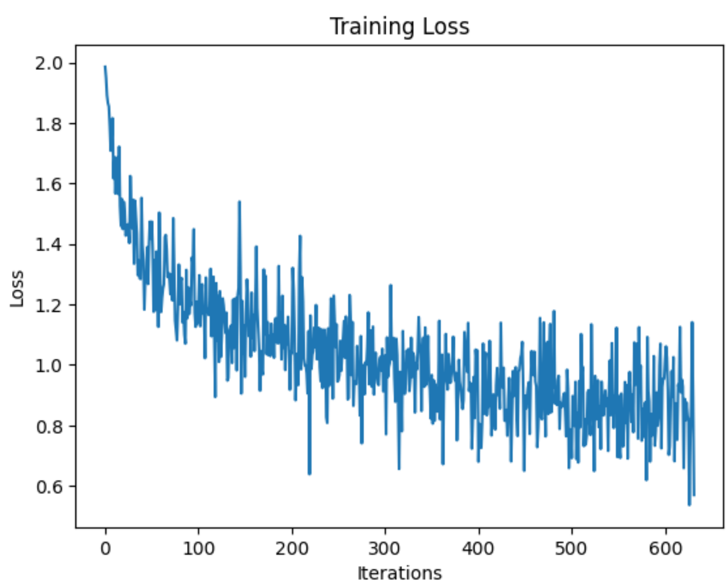

Figure 13: Chart of primary model training loss over iterations

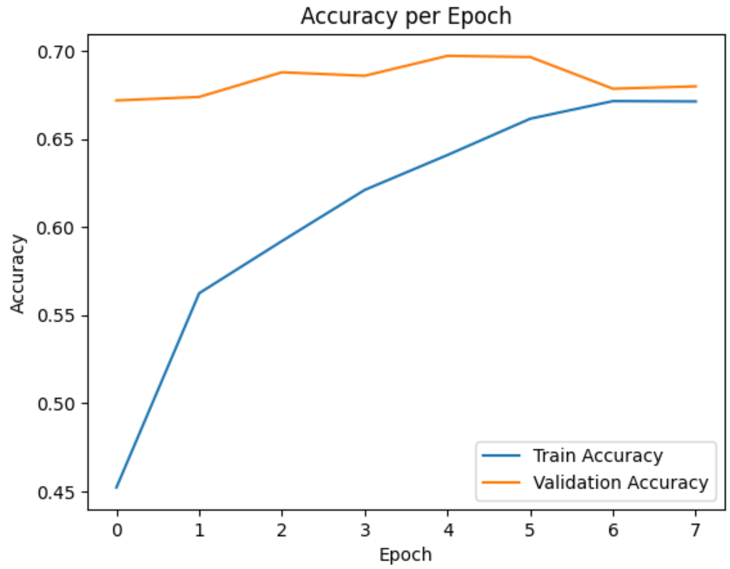

Figure 14: Chart of training and validation accuracy over each epoch

```
Final Training Accuracy: 0.6712
Best Validation Accuracy: 0.6971
```

Figure 15: Final result of current initial primary model

Additionally from a Grad-CAM heatmap we were able to determine that our model was successfully extracting important features from the original image as seen in Figure 16.

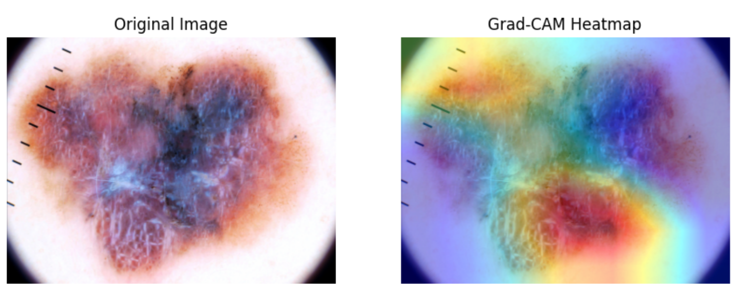

Figure 16: Grad-CAM Heatmap displaying highest classification-contributing geometric features

While these results have shown some improvement from our baseline it still faced very similar problems to our baseline model. The large number of 42000 images in our total training dataset was using a high amount of computational power and training time despite our use of resNet18 and as such we had to limit our training dataset to 5000 images. We also ran into the same problem of accurately predicting the nv class but struggling to predict the other classes. This again was likely due to the large number of nv classes relative to other classes within the validation dataset.

Regardless of this we still have hope of further improving our model, as firstly we are yet to tune any hyperparameters. To do so, we are planning to implement the naive bayes algorithm to find the optimal hyperparameters. During which we will look to either locally train or use stronger GPUs at the UofT computer labs to train our model and do so for longer in order to fully utilize our dataset.

REFERENCES

Canadian Skin Cancer Foundation. Skin cancer facts, 2025. URL https://www.canadianskincancerfoundation.com/skin-cancer/.

Skin Cancer Foundation. Skin cancer facts, 2025. URL https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/.