SKIN CANCER CLASSIFICATION USING DEEP LEARNING MODELS FINAL 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 final project report for Group 29's APS360 final project on classifying various types of skin lesions using deep learning. The report outlines the performance of our baseline and primary models, including implementation details, performance results, and encountered challenges. We describe the use of data preprocessing methods such as data augmentation and optimization algorithms such as random sampling for hyperparameter tuning. Quantitative and qualitative results are provided to demonstrate the feasibility of our approach. The report also includes updated model performance statistics, individual team member contributions, and a discussion reflecting on our project. —-Total Pages: 9

1 INTRODUCTION

Skin cancer is one of the most common malignancies in the world, with approximately 80,000 cases diagnosed annually in Canada Canadian Skin Cancer Foundation (2025). Early and accurate skin cancer recognition significantly improves patient outcomes, yielding a five-year survival rate of 99% Skin Cancer Foundation (2025). However, limited dermatologist access, particularly in rural regions, highlights a persistent gap in timely diagnosis. Furthermore, the most common method of visual examination by dermatologists yields an accuracy of approximately 60% Marks (2002). In this context, ML-based systems hold significant potential to support dermatologists in the diagnostic process for skin cancer Pathan et al. (2022). This project proposes a deep-learning model for classifying dermoscopic images sourced from the HAM10000 dataset into seven skin lesion categories Codella et al. (2021). Deep learning, specifically convolutional neural networks (CNNs), are well-suited due to their ability to learn complex and hierarchical representations directly from raw images. Compared to other traditional models, such as multinomial logistic regression, CNNs are a justified selection for this visual recognition task because their layered architecture enables them to capture both low-level and high-level visual features, such as colour variation and border irregularity which are prominent visual cues commonly used in dermatological diagnosis Al-Antari et al. (2024). Ultimately, this project is a foundational step toward developing end-to-end deployable applications that can be integrated into clinical workflows to assist dermatologists in improving diagnostic process efficiency.

2 ILLUSTRATION

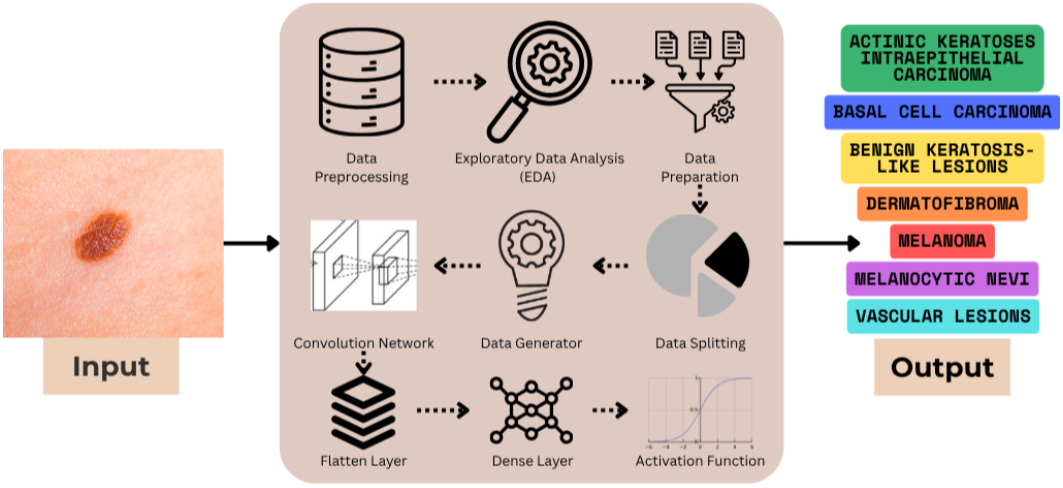

Figure 1: Flowchart displaying our Project's pipeline

3 BACKGROUND & RELATED WORK

This background section provides context on related works that influenced our project's pipeline and architecture.

3.1 ENHANCED SKIN CANCER DIAGNOSIS USING OPTIMIZED CNN ARCHITECTURE AND CHECKPOINTS FOR AUTOMATED DERMATOLOGICAL LESION CLASSIFICATION

This recent study proposed a hybrid classification framework integrating CNNs with XGBoost for classifying dermoscopic images from the HAM10000 dataset. The data preprocessing and cleaning involved artifact removal and normalization of the images. The hybrid model achieved an accuracy of 97.86%, precision and recall of 97.9%, and an F1 score of 97.8%, demonstrating strong performance across multiple classification metrics. The authors concluded that while their model outperformed traditional approaches, the limited size and diversity of the dataset used limited how well the model generalizes to different populations Al-Antari et al. (2024).

3.2 SKIN LESION CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK WITH NOVEL REGULARIZER

This paper proposed a CNN binary classifier to differentiate between malignant and benign skin lesions using the ISIC dataset. The architecture consisted of two convolutional layers followed by pooling, dropout, flattening, and a fully connected layer. A novel regularization method based on the weight matrix's standard deviation was applied to improve generalization, with an optimal regularization parameter of 0.02. The model achieved a maximum average accuracy of 97.49% and demonstrated strong classification ability with good AUC scores. The authors identified the computationally intensive tuning process for the regularization parameter as a limitation in optimization Harangi (2019).

3.3 SKIN LESION CLASSIFICATION USING HYBRID DEEP NEURAL NETWORKS

This study classified skin lesions by combining deep features extracted from AlexNet, VGG16, and ResNet-18, employing an SVM classifier on the ISIC 2017 validation set. This approach yielded AUCs of 83.83% for melanoma and 97.55% for seborrheic keratosis, outperforming individual models and demonstrating improved robustness through the integration of complementary representations across different CNN architectures. Future work was suggested to explore deeper CNNs like DenseNet and the use of localized image patches Esteva et al. (2017).

3.4 A COMPREHENSIVE STUDY ON SKIN CANCER DETECTION USING ARTIFICIAL NEURAL NETWORK (ANN) AND CONVOLUTIONAL NEURAL NETWORK (CNN)

This study emphasized the superiority of CNNs over traditional ANNs in identifying features critical to skin lesion classification. A hybrid system combining CNNs with NLP achieved 99.35% on training accuracy; however, test accuracy ranged between 66–83%, revealing robustness concerns. The authors highlighted key challenges pertaining to the datasets being inclusive of diverse lesion types and skin tones Brinker et al. (2019).

3.5 AUTOMATED SKIN LESION CLASSIFICATION USING ENSEMBLE OF DEEP NEURAL NETWORKS IN ISIC 2018: SKIN LESION ANALYSIS TOWARDS MELANOMA DETECTION CHALLENGE

This paper evaluated several advanced CNN architectures on the ISIC 2018 dataset, with PNASNet-5-Large achieving a notable validation score of 0.76. The study emphasized the efficiency of CNNs in skin lesion classification while noting the role of hyperparameters optimization and diverse datasets in improving the generalizability of these models Tajbakhsh et al. (2024).

4 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 2

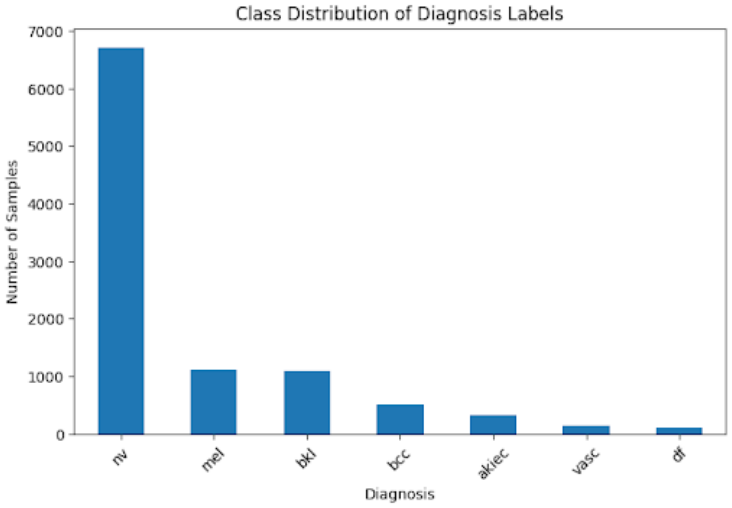

Figure 2: Image displaying the sample distribution amongst classes

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

4.1 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 3.

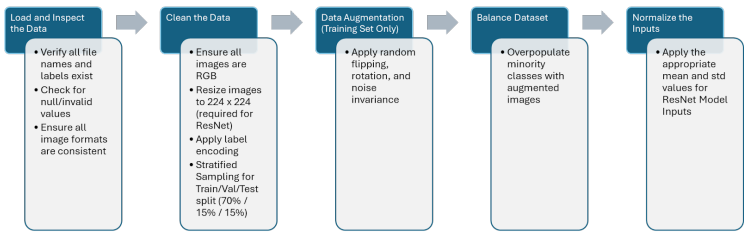

Figure 3: Display of data cleaning and preprocessing process

4.2 Plan for Testing Data

For final evaluation, we tested 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.

4.3 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 underrep-

resented classes like melanoma. We addressed this through oversampling the minority classes via data augmentation.

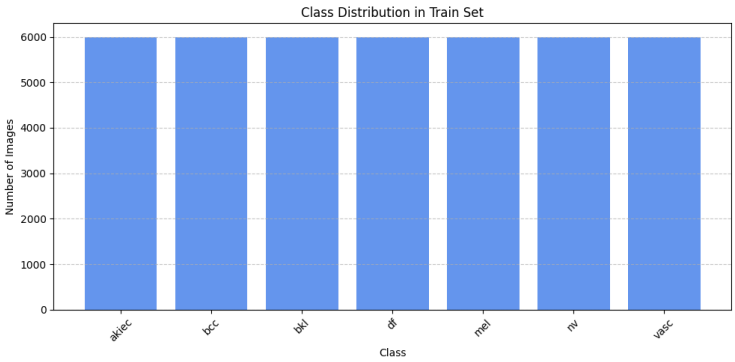

Figure 4: Display of the balanced class distribution following oversampling via data augmentation

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.

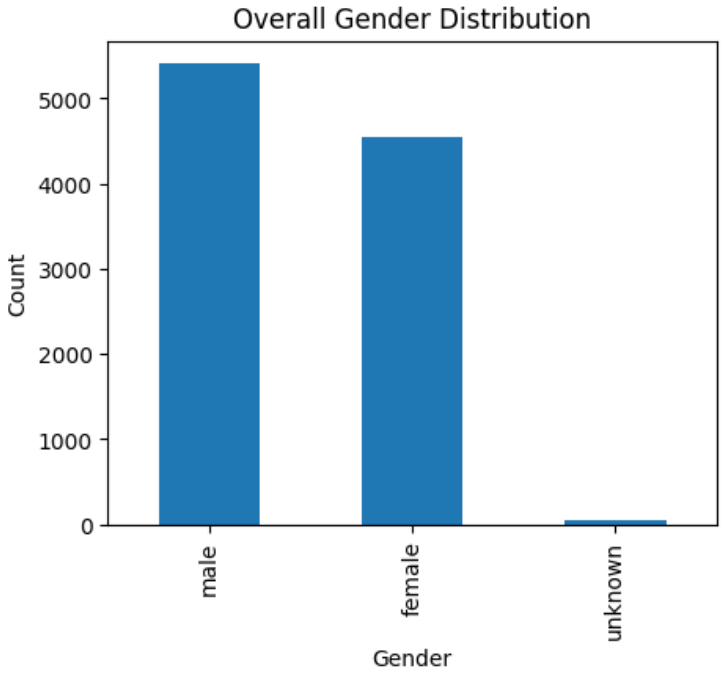

Figure 5: Distribution of dataset samples across genders

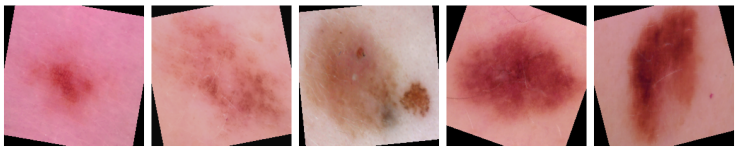

Figure 6: Five image samples following data augmentation

5 ARCHITECTURE

We used a pre-trained ResNet-18 convolutional neural network (CNN) as our final architecture. ResNet-18 is a residual network with 18 layers which was made to fix the vanishing gradient problem in deep networks by introducing skip connections (also known as residual connections) He et al. (2015). These connections allow the network to learn identity mappings which allows deeper architectures to train properly without degradation in performance.

The encoder portion of ResNet-18 consists of an initial 7×7 convolutional layer with stride 2, followed by batch normalization, a ReLU activation, and a 3×3 max pooling layer. This is followed by four sequential residual blocks, each containing two 3×3 convolutional layers and a skip connection that adds the input of the block to its output. As we move deeper into the network, the number of filters progressively increases from 64 to 512, allowing the model to capture progressively more complex and abstract visual features such as texture patterns and color irregularities in skin lesions.

The decoder portion of ResNet-18 uses a global average pooling layer, which reduces the spatial dimensions of the final feature maps, followed by a fully connected layer. In our case, we have replaced the original fully connected layer with one which has seven output neurons, corresponding to the lesion categories. A softmax activation is then applied to this layer to produce a probability distribution over the classes.

The model uses the CrossEntropyLoss function, and its parameters were updated through backpropagation with the Adam optimizer for efficient convergence. To improve performance and generalization, we then tuned key hyperparameters such as learning rate and weight decay.

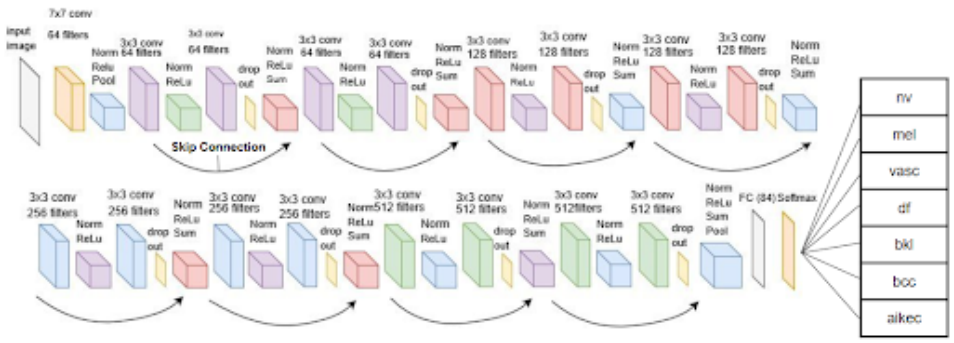

Figure 7: Display of our deep learning model's architecture

6 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.

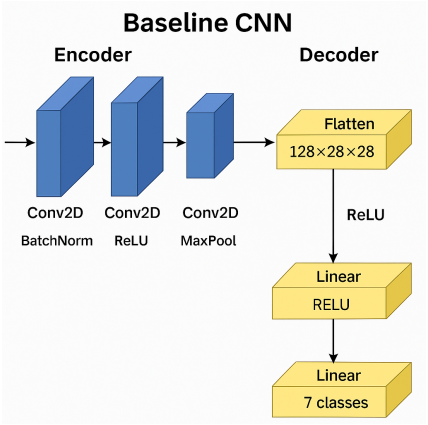

Figure 8: Baseline model architecture

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

7 Quantitative Results

We evaluated our model using accuracy as the primary metric, as it provides a straightforward measure of the proportion of correctly classified skin lesion images. This metric was chosen because the dataset, after oversampling, had balanced class representation, making accuracy an appropriate and interpretable measure. Prior to hyperparameter tuning, the primary model achieved 67.12% training accuracy and 69.71% validation accuracy as seen in Figure 3 and Figure 4. After hyperparameter tuning (learning rate and weight decay factor), the model achieved 75.6% validation accuracy as seen in Figure 11. These results indicate that leveraging pretrained networks and fine-tuning parameters improved the model's ability to generalize to unseen data.

```
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 9: Display of our primary model's train and validation accuracy metrics

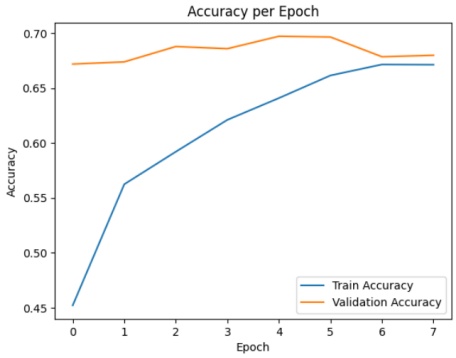

Figure 10: Chart of training and validation accuracy over each epoch for primary model

```
Final Training Accuracy: 0.8260
Best Validation Accuracy: 0.7800
```

Figure 11: Results from random sampling for hyperparameter tuning

8 Qualitative Results

A prediction gallery from validation split without augmentation is shown in Figure 12. Each tile is clearly annotated with the predicted and true label for clear visualization. A clear pattern in the gallery is the model's reliability on nevus (nv). Because the original training set is heavily skewed toward nv, and the validation split also contains many nv cases. The model has learned nv appearance well and often predicts it correctly, especially when lesions are centered, show relatively uniform pigmentation, and have clean borders. In contrast, performance is less stable on non-nv classes, and are more frequently misclassified. Visually, these mistakes tend to occur when lesions are small, low-contrast, off-center, or partially obscured by hair or glare. Overall, the gallery introduces a class-imbalance effect, with strong nv recognition, coupled with a bias toward nv on ambiguous or underrepresented cases.

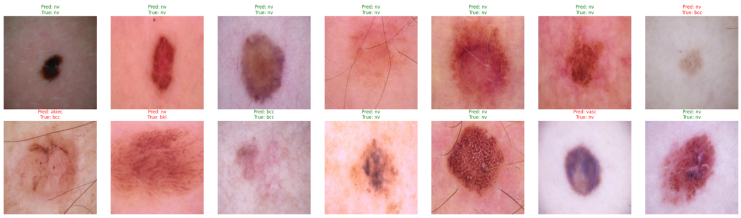

Figure 12: Prediction gallery displaying predictions and true labels on different types of skin cancer lesions

Several Grad-CAM examples are shown in figure 13, computed on the last layer of the model. In this image, the heatmap focuses on the lesion body and border network rather than background skin, indicating that the model is attending to clinically meaningful cues (texture, pigment network, and edge structure). While these examples are not exhaustive, it is consistent with the gallery's high-confidence nv predictions: the network is "looking in the right place." At the same time, the class distribution likely shapes the final decision. Our original training set prior to any rebalancing, and our validation split are nv-heavy. By design, we did not augment validation to keep it a clean, unbiased estimate of generalization. This means both the training signal and the evaluation feedback emphasize nv, which could be a cause for the model predicting nv on borderline cases even when its attention is lesion-centric.

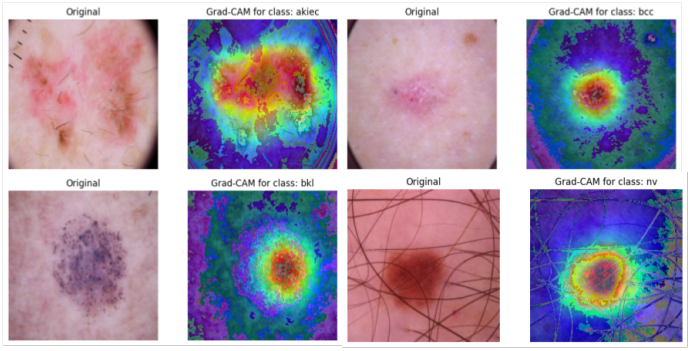

Figure 13: Grad-CAM heatmap displaying which geometric features the model focuses on

The results shown in the images suggest the model is learning meaningful, lesion-centric cues, border structure and internal texture when cases are clear, which matches the high-confidence correct

nv examples. Errors appear more often on rarer or visually ambiguous cases and when hair or glare pulls attention outside the lesion. Class imbalance likely contributes to these confusions, but it isn't the only factor. As next steps, we'll apply modest class balancing and minority-focused augmentation, try higher-resolution or lesion-centered crops to preserve border detail, and add simple artifact handling to improve consistency without changing the architecture.

9 Evaluating model on new data

Due to limitations with the nature of the project, we were not able to recreate unbiased images of skin cancer lesions. As such, we evaluated our model on a completely held-out test set, 15% of the HAM10000 dataset, consisting of skin lesion images that were never used during training, validation, or influence in tuning our hyperparameters. This ensured an unbiased estimate of real-world performance. The model achieved high accuracy of 70.5% demonstrating strong performance to unseen data. Additionally we found images of skin cancer lesions from Google searches which were not a part of the HAM10000 dataset and manually tested them on the model. This yielded a similar result to before of approximately 70% accuracy further validating the performance from before.

10 Discussion

Our results demonstrate that the ResNet18 transfer learning model outperformed the baseline CNN in both accuracy and efficiency, confirming the effectiveness of pre-trained features. During testing, the primary model demonstrates strong capability in identifying most skin lesion classes with particularly high accuracy, however sometimes struggled to distinguish between Basal Cell Carcinoma (BCC) and Benign Keratosis-like Lesions (BKL), likely due to their visual similarity and overlapping features in the dataset. Grad-CAM visualizations confirmed that both models often focused on medically relevant lesion areas, but ResNet18 showed more consistent and localized attention patterns. Additionally, through testing the model on a few images extracted from websites such as National Cancer Institute (NIH), it was able to appropriately classify most images demonstrating that the model's development was successful.

A major limitation of this project was GPU resources, as we were limited to the provided GPU space from google colab which was not large enough to allow advanced search algorithms like bayesian search on a reasonable amount of the dataset. Instead we had to resort to Random Sampling to optimize our hyperparameters. To offset this loss of performance and complexity, we implemented a two phase training scheme. We first froze the convolutional base and ran a couple epochs in order to extract features. Then unfroze everything and continued training on a lower learning rate to fine tune the model Chollet (2020). The effects of this can be found from our Grad-CAMs.

11 Ethical Considerations

Skin cancer classification could present several ethical challenges, particularly around fairness and potential harm from misclassification. A major concern is that commonly used datasets, such as the HAM10000, lack diversity in skin tone representation, with very few examples of dark skinned patients Morales-Forero et al. (2024). This may result in a model that is poor in classifying underrepresented groups, thereby inflicting the risk of misclassification. A well-trained model will not always provide accurate results. Identifying the wrong skin cancer type could be harmful for patients, as an incorrect diagnosis could result in unnecessary treatment, or failure to treat a serious condition in time. These challenges highlight the importance of evaluating model performance across diverse groups, and clearly communicating the model's limitations when interpreting results.

REFERENCES

Mohammad A. Al-Antari et al. Explainable deep learning model for melanoma classification using dermoscopy images. *BMC Medical Imaging*, 24(1), 2024. doi: 10.1186/s12880-024-01356-8. URL https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-024-01356-8.

Titus Josef Brinker et al. Skin cancer classification using convolutional neural networks: systematic review, 2019. URL https://arxiv.org/pdf/1901.10802.

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

François Chollet. Transfer learning & fine-tuning. Online; last modified 2023-06-25, 2020. URL https://keras.io/guides/transfer_learning/.

Noel C. Codella et al. Skin lesion analysis towards melanoma detection, 2021. URL https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T. Harvard Dataverse.

Andre Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks, 2017. URL https://arxiv.org/pdf/1702.08434.

Balazs Harangi. Skin lesion classification with ensembles of deep convolutional neural networks. *Journal of Biomedical and Health Informatics*, 2019. doi: 10.1109/JBHI.2019.2899327. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8669763.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. URL https://arxiv.org/abs/1512.03385.

R. Marks. An overview of skin cancers: incidence and causation. *The Lancet Oncology*, 3(4):165–169, 2002. doi: 10.1016/S1470-2045(02)00679-4. URL https://www.thelancet.com/journals/lanonc/article/PIIS1470204502006794/abstract.

Andres Morales-Forero, Lili Rueda Jaime, Sebastian Ramiro Gil-Quiñones, Marlon Y. Barrera Montañez, Samuel Bassetto, and Eric Coatanea. An insight into racial bias in dermoscopy repositories: A ham10000 data set analysis. *JEADV Clinical Practice*, 3(3):836–843, 2024. doi: 10.1002/jvc2.477. URL https://onlinelibrary.wiley.com/doi/full/10.1002/jvc2.477.

Shakir Pathan et al. Melanoma skin cancer detection using deep learning and cnn. *Journal of Healthcare Engineering*, 2022. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9693628/.

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

Nima Tajbakhsh et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? *Journal of Visual Communication and Image Representation*, 2024. doi: 10.1002/jvc2.477. URL https://onlinelibrary.wiley.com/doi/full/10.1002/jvc2.477.