Lesion Classification: Malignant or Benign

Andre Menezes, Tony Lin, William Clubine
{meneza3,lint50,clubinew}@mcmaster.ca

1 Introduction

Skin cancer is one of the most common malignancies worldwide, and is becoming even more common over time (Karp et al., 2024). Malignant skin cancers are more commonly known as melanomas, and although this subtype makes up only about 1% of all skin tumours, it accounts for the vast majority of skin-cancer deaths. This means that early detection is critical, as the five-year survival is very high when caught early (in the 90–99% range), but drops sharply if diagnosis is delayed (Naqvi et al., 2023). In Canada and the U.S., roughly one in five people will develop some form of skin cancer by age 70. This makes automated diagnostic tools appealing for assisting dermatologists and patients.

In this project, we aim to build a machine-learning image classification model to label lesions as malignant or benign. By training on dermoscopic images, the model can capture subtle visual patterns and help flag suspicious lesions. We do not expect perfect accuracy, as expert dermoscopic exams are only about 60% accurate in practice (Al-Waisy et al., 2025), but a high-performing algorithm could make early screening significantly faster and more widely accessible. Skin-cancer classification is also inherently difficult due to high visual similarity between lesion types and large variability within each class (Wu et al., 2022).

Recent research has extensively applied deep learning to skin-lesion classification. Convolutional neural networks (CNNs) can learn rich visual features directly from images, and many studies report very high accuracy on standard datasets. For example, Esteva et al. (2017) trained a CNN on more than 100,000 clinical images and demonstrated dermatologist-level performance on malignant vs. benign classification. Others have fine-tuned pretrained models such as AlexNet, VGG, ResNet, DenseNet, and NASNet on dermoscopy datasets—including the one we use in our project—and often achieved classification accuracies above 95%. However, most dermatology datasets remain small and highly imbalanced, which limits robustness and real-world generalization (Wu et al., 2022). One study found that a DenseNet-121 feature extractor plus a neural classifier reached approximately 98.3% accuracy (Naqvi et al., 2023), and ensemble or attention-based architectures such as "Skin-DeepNet" have pushed reported accuracy even higher (Al-Waisy et al., 2025).

Despite this progress, high performance on curated dermoscopic images does not guarantee reliability in uncontrolled settings. In practical deployment, many users capture lesions with smartphone cameras rather than clinical dermatoscopes. These images vary widely in focus, lighting, distance, and framing, producing a significant domain shift that dermoscopy-trained models often fail to handle. Clinical evidence shows that patient-captured mobile-phone images are frequently of insufficient quality for diagnosis, with inadequate information in roughly one-third of cases (Weingast et al., 2013). This highlights a key gap between benchmark results and real-world usage.

Beyond binary malignant–benign classification, more recent work has explored multi-modal systems that incorporate metadata such as patient age, lesion location, and clinical history. For example, Ahmadi Mehr and Ameri (2022) achieved approximately 89.3% accuracy when jointly modeling images and patient attributes to classify four common skin conditions. We build on this foundation by developing our own CNN-based classifier with the goal of achieving strong diagnostic accuracy for binary lesion classification while following best practices for transfer learning and model evaluation.

After reviewing related work, particularly studies highlighting the performance gap between dermoscopic datasets and real-world smartphone im-

ages, we refined our understanding of the domain-shift challenges in skin-cancer classification. While this did not change our core project direction, it informed our expectations regarding model generalization.

2 Dataset

The dataset we were originally using can be found on Kaggle and consists of 3297 images of skin lesions. Approximately half of these (1800) show benign skin lesions, and the other half (1497) show malignant skin lesions, which is annotated in the dataset by splitting the images into two distinct folders based on their classification.

These images were originally sourced from the International Skin Imaging Collaboration (ISIC), which provides an open archive of thousands of images captured and hand-annotated by clinicians all over the world. This means that the annotations are based on diagnoses from trained healthcare professionals, and thus we can expect them to be accurate with a relatively high degree of certainty.

2.1 Dataset Changes

We switched to the 10,000-image ISIC dataset on Kaggle to address limitations we saw when training on the smaller set. This dataset consists of 5500 images of benign skin lesions and 5105 images of malignant skin lesions. In our previous dataset with fewer samples, the model tended to overfit and showed weaker generalization on the validation split, or under fit if data augmentation was applied to address the overfitting. This larger dataset provides a wider range of lesion appearances and natural variation across patients, which supports more stable learning and reduces the model's dependence on augmentation alone. Since the new dataset follows the same ISIC format and class structure, our pre-processing steps were mostly unchanged, aside from a few extra augmentation steps including perspective shifts and ImageNet auto-augmenting. The existing normalization, cropping, and augmentation pipeline still applies directly and continues to serve the same purpose of improving generalization. The main practical difference introduced by the larger dataset is increased training and data-loading time, but the overall workflow remains unchanged.

2.2 Preprocessing

To prepare the images for the model, each image is first normalized by the ImageNet mean and standard deviation, and resized to a standard resolution of 300×300 pixels to match the expected input of the backbone model (see implementation details). Additionally, for the images used in training the model, we apply random crops, flips, rotations, perspective shifts, and color adjustments to simulate variations in orientation, lighting, and framing that may occur in real-world images. These augmentations help the model generalize better by preventing overfitting to specific spatial or color patterns (Shorten and Khoshgoftaar, 2019). Random erasing is also used as a regularization technique, forcing the model to rely on broader contextual features rather than any single localized region.

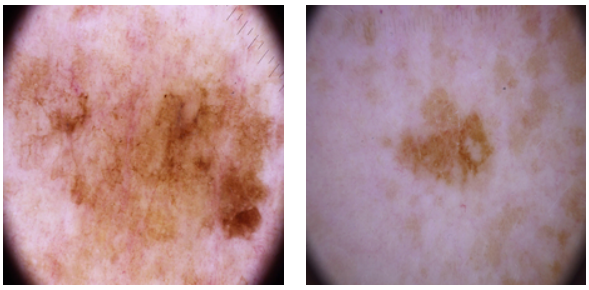

Figure 1: A malignant (left) and benign (right) lesion.

3 Features

Following prior work in deep learning-based skin lesion analysis (Esteva et al., 2017; Naqvi et al., 2023; Al-Waisy et al., 2025), our model relies on a convolutional neural network (CNN) to learn visual features directly from raw dermoscopic images rather than through handcrafted feature extraction. Each image of a skin lesion is sized to 300×300 pixels and normalized using the standard ImageNet mean and standard deviation so that its pixel distribution aligns with the pretrained network weights. We do not incorporate any additional metadata (such as age, sex, or lesion location), so all information used by the model comes from the dermoscopic images themselves.

The feature extraction process is handled by a EfficientNetB3 architecture pretrained on ImageNet (ImageNet1K_V2), used here as a transfer learning backbone. This network serves as a hierarchical feature extractor: earlier convolutional layers detect low-level visual cues such as edges, colour gradients, and small texture variations, while deeper layers encode higher-level, semantically rich patterns related to lesion structure, border irregularities, and shape asymmetries. These progressively abstracted representa-

tions form a 2048-dimensional embedding vector that summarizes each image in a compact feature space. Using a pretrained backbone makes sense in our setting because the dataset is relatively small; transferring generic visual filters from ImageNet reduces overfitting and has been repeatedly shown to improve performance on medical images (Esteva et al., 2017).

On top of this backbone, we replace the original 1000-class output layer of EfficientNetB3 with a new two-unit linear head corresponding to benign and malignant classes. This classifier operates on the 2048-dimensional embeddings produced by the backbone and maps them to softmax probabilities for binary prediction. During training, the model jointly optimizes these representations so that embeddings of malignant and benign samples become separable in feature space. In doing so, the system learns the relevant visual characteristics for diagnosis directly from data, rather than depending on predefined colour or texture descriptors.

We do not perform any explicit feature selection: all channels in the dermoscopic images are kept, and the full 2048-dimensional embedding from the backbone is passed to the classifier without manual pruning. Instead of dropping or hand-picking features, we rely on the network's internal representation learning and regularization to suppress uninformative patterns.

However, we do change the *effective* input distribution through data augmentation. For training images, we apply random crops, horizontal flips, rotations, colour jitter, perspective shifts, and random erasing as part of the preprocessing pipeline. These augmentations were added because they simulate realistic variation in orientation, lighting, framing, and partial occlusion, which should encourage the model to focus on lesion shape and structure rather than memorizing specific pixel arrangements.

In terms of varying features across experiments, our main comparison was between using the ImageNet-pretrained EfficientNetB3 as a frozen feature extractor with only the final linear layer trained, versus fine-tuning the entire backbone end-to-end. The frozen variant relies on generic natural-image features, while fine-tuning adapts the representation to dermoscopic patterns observed in our dataset.

4 Implementation

As described in the Features section, the final model is built on an EfficientNet-B3 convolutional backbone, which provides a favourable balance between representational capacity and computational efficiency. The implementation adapts this pretrained architecture to the binary task of benign vs malignant skin lesion classification and integrates it into a training pipeline. All components are implemented in PyTorch, which enables modular design, scalable experimentation, and flexible optimization.

The network is initialized with ImageNet-pretrained weights (ImageNet1K) to leverage generalized visual features such as colour gradients, texture patterns, and edge structures learned from large-scale natural image data. To specialize the model for skin lesion classification, the original classification head is replaced with a new fully connected layer that outputs two logits corresponding to the target classes. This layer maps the 1536-dimensional pooled feature vector from the EfficientNet backbone to the final binary prediction. Only this new layer is randomly initialized. All remaining parameters retain their pretrained weights.

In order to achieve stable convergence, training is organized into two distinct stages. During the initial warmup phase (3 epochs), the backbone is frozen, and only the classification head is trained. This allows the new head to learn class-specific decision boundaries without affecting the pretrained feature representations. The optimizer used is AdamW, with a learning rate of 5×10^{-4} for the head and a weight decay of 5×10^{-5}.

After warmup, the model enters the fine-tuning stage, in which the entire network is unfrozen and trained end-to-end for up to 50 epochs, with early stopping applied if validation performance fails to improve by at least 1×10^{-4} over five consecutive epochs. In practice, training typically converges after approximately 30 epochs, though this does vary. A differential learning rate strategy is used, applying a lower rate (1×10^{-4}) to the pretrained backbone while maintaining 5×10^{-4} for the classification head. A cosine annealing learning rate scheduler gradually reduces the learning rate across epochs, improving generalization and preventing overfitting.

The model is trained using the cross-entropy loss function with a batch size of 256. Automatic mixed-precision (AMP) is enabled on GPU to accelerate

training and reduce memory usage. Each epoch consists of forward propagation, loss computation, backpropagation, and parameter updates, followed by validation to track performance. To ensure reproducibility, random seeds are fixed, and model checkpoints are saved whenever validation metrics improve.

Data loading and preprocessing are handled using PyTorch's ImageFolder and DataLoader APIs. Separate loaders are defined for training, validation, and test splits. Training images go through extensive on-the-fly augmentation through torchvision transforms, including random cropping, flipping, rotation, color jitter, random erasing, perspective distortion, simulated lighting variation, and ImageNet AutoAugment policies. These augmentations are used to improve robustness by simulating real-world variability and preventing overfitting. Validation and test images are processed with deterministic resizing and center cropping to ensure consistent and unbiased evaluation.

After training concludes, the model's output probabilities are post-processed using threshold tuning. Rather than applying a fixed threshold of 0.5, the decision threshold is selected to maximize validation-set performance, optimizing either accuracy or F1-score depending on the evaluation target. This final calibration step improves classification reliability, particularly in the presence of class imbalance and probabilistic uncertainty.

The implementation follows a modular and extensible design. Separate modules manage data handling, model definition, optimization, evaluation, and checkpointing. This architecture makes it straightforward to substitute alternative backbones, adjust hyperparameters, or introduce new training strategies without modifying the core pipeline. All experimental outputs, including trained weights, logs, and configuration files are stored in a checkpoints directory to ensure traceability and reproducibility.

5 Results and Evaluation

To evaluate our model's performance in distinguishing malignant vs. benign skin lesions, we used a structured two-phase training process with separate warm-up and fine-tuning stages. During the warm-up phase, only the classifier head was trained on frozen ImageNet-pretrained backbones (EfficientNetB3), ensuring that the model adapted its higher-level features to the new dataset while avoiding overfitting. The fine-tuning phase then unfroze the backbone for end-to-end optimization at a lower learning rate, improving representational alignment with the medical domain.

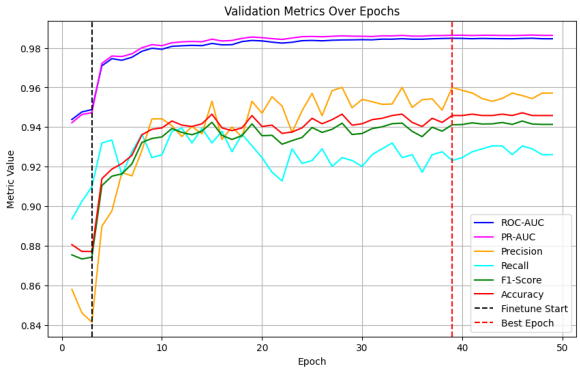

Figure 2: Validation performance through epochs.

The primary metric for model selection was the ROC-AUC (Receiver Operating Characteristic - Area Under Curve), which measures the model's ability to distinguish between malignant and benign samples regardless of threshold. In addition to that, some performance indicators (see Figure 2) include Precision, Recall, F1-score, PR-AUC (Precision-Recall AUC), and Accuracy, along with a confusion matrix for class-wise inspection (see Figure 4). All these metrics were computed using scikit-learn functions.

As a baseline, we compared against a non-finetuned ImageNet model, which achieves near-random classification performance (ROC-AUC $\approx$ 0.5) when directly applied to the dataset. After warm-up and fine-tuning, our model showed a significantly higher ability to distinguish between the two classes, as observed by validation ROC-AUC values reaching 0.985 during training (see Figure 2). The consistent upward trend in both ROC-AUC and PR-AUC across epochs indicated stable learning and generalization improvements. Using these results, the best-performing checkpoint (highest validation ROC-AUC) was saved for later deployment and testing.

Ultimately, we obtained a model that achieved an accuracy of 97.15% on the training set, 95.49% accuracy on the validation set (see Figure 3), and 93.30% accuracy on the test set (see Figure 4). These evaluation results demonstrate that transfer learning, even with limited labeled dermatological data, can achieve high diagnostic accuracy when paired with careful data augmentation, validation-based early selection, and multi-metric

Figure 3: Training/validation accuracy through epochs.

performance tracking.

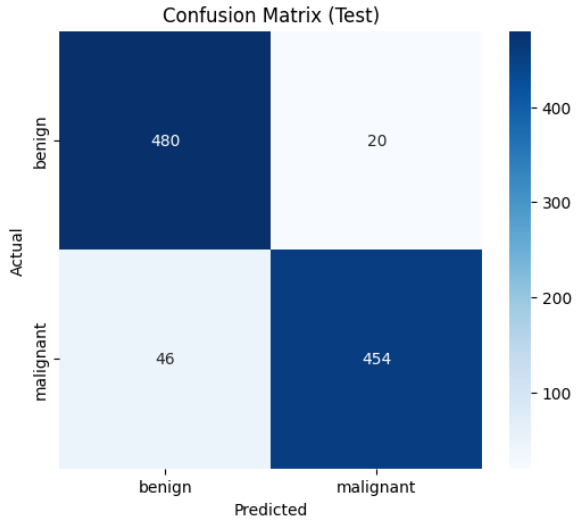

Figure 4: Confusion matrix after training.

6 Progress

The development of the final model mostly followed the plan outlined in the progress report, with several planned improvements successfully implemented and additional refinements introduced during experimentation. The objectives post-progress report focused on improving generalization through data augmentation, increasing training efficiency through architectural changes, and adding early stopping to reduce overfitting and unnecessary computation.

All major planned changes were completed. Data augmentation was substantially expanded beyond basic transformations to include domain-specific variations, improving the robustness of the model to real-world image conditions. Weighted sampling was applied to the dataset to ensure class imbalances (however minimal they may be) are counteracted. The backbone architecture was successfully upgraded from ResNet-50 to EfficientNet-B3, as originally proposed, enhancing efficiency and representational power without sacrificing performance. Early stopping was also implemented as planned, enabling training to terminate automatically once validation performance plateaued, which reduced training time and mitigated overfitting. And the entire process was made easier thanks to a fully refactored code base in a Jupyter notebook, replacing the original python file and allowing us to iterate without re-running the entire process every time.

While the majority of the original plan was followed, some components evolved during the course of development. Fine-tuning was extended from the originally planned schedule to a longer training regime supported by stronger regularization and early stopping. In addition, decision threshold optimization was introduced after training to improve final classification performance, particularly with respect to accuracy and F1-score. This calibration step was not part of the initial plan but was added after observing that probability outputs could be further optimized beyond a fixed threshold.

Overall, the project remained aligned with its original design goals while evolving through experimental findings and performance evaluation. These changes reflect an iterative development process, resulting in a final model that is more robust, efficient, and clinically relevant than originally planned.

7 Error Analysis

To examine our model's errors systematically, we tracked threshold-independent metrics during training (ROC-AUC and PR-AUC) alongside threshold-dependent metrics (accuracy, precision, recall, and F1) and visual diagnostics. Epoch-by-epoch plots of these metrics (see Figure 2) allowed us to distinguish score quality from classification calibration. ROC-AUC/PR-AUC indicated how well the model separates benign vs malignant cases across thresholds, while accuracy/F1 reflected our specific operating point (initially at 0.5, later tuned on validation). We complemented these with confusion matrices on the test set (see Figure 4), which made false positives and false negatives visible at a glance. Precision–recall and ROC curves (see Figure 5) help us to further understand performance under class imbalance, showing how recall improves

as the threshold drops while precision degrades.

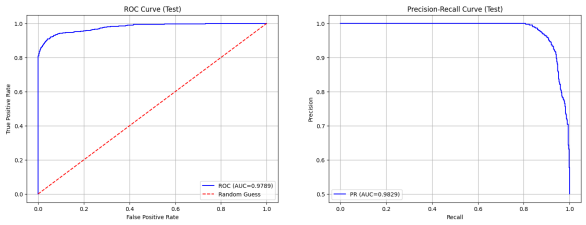

Figure 5: ROC and Precision-Recall curves.

By examining the samples that our model fails to classify correctly (e.g. by looking at those in the false positive and false negative regions of the confusion matrix), we were able to develop a better understanding of what cases our model does not handle well. Misclassified benign images often exhibited atypical texture, lighting, or lesions resembling malignant patterns, while malignant misses tended to be small, low-contrast lesions or cases with occlusion and challenging illumination. Improving our models handling of these cases while preserving its performance on already correct cases proved challenging to balance. For example, weighted sampling slightly improved the recall for malignant cases but sometimes increased false positives in benign, reflected by the confusion matrices shifting off-diagonal.

Overall, the model appears good at clear, well-lit lesions with consistent backgrounds. It is weaker on borderline cases or photographs with poor quality (strong shadows, low contrast, occlusions, or unusual artifacts). If we were to continue to work on this model, we would try to address these issues by further tuning our data augmentation to improve generalization while keeping accuracy high, and by experimenting with techniques like supervised attention in order to address cases where lesions are smaller or low-contrast. We may also try incorporating additional metadata such as patient age, gender, race, and existing conditions in order to reduce that impact of image quality issues or irregularities. This would be relatively easy to obtain, as most skin lesion datasets already include this additional metadata. In general, we would continue to focus on examining which cases our model can and cannot handle and adjust our architecture accordingly with further experimentation and comparison.

8 Team Contributions

8.1 Progress Report

Model implementation and training were conducted collaboratively by Andre Menezes and Tony Lin. The dataset was prepared/sourced by William Clubine, who also evaluated the final trained model. We divided the progress report based on these responsibilities, with Andre Menezes writing Results and Evaluation and Feedback and Plans, Tony Lin writing Features and Implementation, and William Clubine writing Introduction, Related Work, and Dataset. All members worked together to review each other's work and prepare the final submission.

8.2 Final Report

Andre Menezes modified Evaluation and Progress. Tony Lin generated the Jupyter notebook and modified the following sections: Introduction, Dataset, Features and Inputs, and Team Contributions. William Clubine refactored, upgraded, and fine tuned the model architecture based on the Progress Report's Feedback and Plans and modified the Implementation and Error Analysis sections. All members worked together to review each other's work and prepare the final submission.

References

Reza Ahmadi Mehr and Ali Ameri. 2022. Skin cancer detection based on deep learning. *Journal of Biomedical Physics and Engineering*, 12(6):559–568.

Alaa S. Al-Waisy, Shumoos Al-Fahdawi, Mohammed I. Khalaf, Mazin Abed Mohammed, Bourair Al-Attar, and Mohammed Nasser Al-Andoli. 2025. A deep learning framework for automated early diagnosis and classification of skin cancer lesions in dermoscopy images. *Scientific Reports*, 15(1):31234.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Paulina Karp, Katarzyna Karp, Marcelina Kądziela, Radosław Zajdel, and Agnieszka Żebrowska. 2024. The importance of early detection and prevention of atypical skin lesions and other melanoma risk factors in a younger population. *Cancers*, 16(24).

Maryam Naqvi, Syed Qasim Gilani, Tehreem Syed, Oge Marques, and Hee-Cheol Kim. 2023. Skin cancer detection using deep learning—a review. *Diagnostics*, 13(11).

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Jessika Weingast, Christian Scheibböck, Elisabeth Wurm, Elisabeth Ranharter, Stefanie Porkert, Stephan Dreiseitl, Christian Posch, and Michael Binder. 2013. A prospective study of mobile phones for dermatology in a clinical setting. *Journal of telemedicine and telecare*, 19:213–8.

Yinhao Wu, Bin Chen, An Zeng, Dan Pan, Ruixuan Wang, and Shen Zhao. 2022. Skin cancer classification with deep learning: A systematic review. *Front Oncol*, 12:893972.