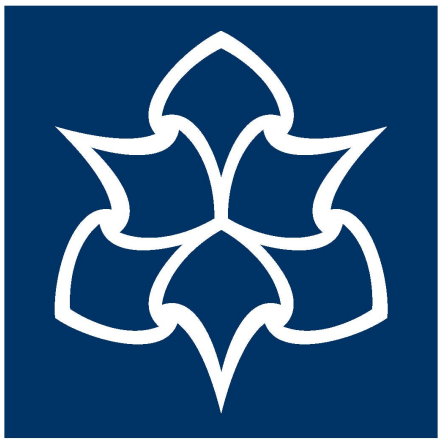
Using ResNet18 to Classify Skin Lesions from Medical Images: A Statistical Learning Approach

A DISSERTATION SUBMITTED TO MANCHESTER METROPOLITAN UNIVERSITY
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE FACULTY OF SCIENCE AND ENGINEERING



2025

By Fundo Lazaro Ambrosio Fernando
Department of Computing and Mathematics

Table of Contents

List of Tables

List of Figures

Abstract

Skin cancer is one of the most common and potentially fatal diseases globally, with early diagnosis being critical for successful treatment. Traditional methods such as dermoscopic examination are prone to variability due to reliance on clinician expertise, especially among less experienced dermatologists. This project explores the application of deep learning to support automated skin lesion classification using dermoscopic images.

The study focuses on fine-tuning a ResNet15 architecture to classify images from the publicly available HAM10000 dataset, which includes seven diagnostic categories. To evaluate comparative performance, DenseNet121 and EfficientNet-B0 were also implemented under identical training conditions. All models were assessed using standard classification metrics including accuracy, precision, recall, and F1-score.

Additionally, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to visualize class-discriminative regions in the images, improving the interpretability of predictions. The ResNet18 model achieved the most balanced performance across all metrics, particularly excelling in identifying minority classes.

The results demonstrate that ResNet18 can serve as an effective backbone for clinical decision support systems in dermatology, offering both high classification performance and visual transparency. The prototype application further illustrates how deep learning models can be made accessible and interpretable in healthcare environments.

Declaration

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work. This work has been carried out in accordance with the Manchester Metropolitan University research ethics procedures, and has received ethical approval number Your EthOS Number.

Signed: Fundo L.A. Fernando

Date:03/09/2025

Acknowledgements

First and foremost, I thank God Almighty for granting me strength, wisdom, and perseverance throughout this journey. Without His guidance, this achievement would not have been possible.

I am deeply grateful to my academic supervisor**,** Dr. Philip Sinclair, for his unwavering support, expertise, and constructive feedback throughout the course of this research. His guidance was invaluable in shaping the direction and depth of this work.

My sincere thanks to Manchester Metropolitan University and the Department of Computing and Mathematics for providing an enriching learning environment and access to the resources that made this project possible.

I also wish to express my heartfelt appreciation to the Chevening Scholarship Programme for funding my postgraduate studies. This opportunity has been a pivotal step in both my academic and professional development.

To my parents, I owe a profound debt of gratitude for their lifelong encouragement, values, and belief in the power of education. Their constant motivation has always pushed me to pursue excellence.

Special thanks to my wife and family for their unconditional love, sacrifices, and continuous support throughout my studies.

Lastly, I acknowledge the creators of the HAM10000 dataset and the open-source communities whose tools and contributions played a vital role in the technical execution of this project.

To all who supported me in this academic journey, emotionally, and spiritually thank you

Abbreviations

CNN Convolutional Neural Network.
AUC Area Under the Curve.
AI Artificial Intelligence.
GPU Graphics Processing Unit.
HAM10000 Human Against Machine with 10000 training images.
ReLU Rectified Linear Unit.
ILSVRC ImageNet Large Scale Visual Recognition Challenge.
AKIEC Actinic Keratoses and Intraepithelial Carcinoma.
MEL Melanoma.
CT Computed Tomography.
LIDC-IDRI Lung Image Database Consortium and Image Database Resource Initiative.
LTA Long-Term Agreement.
ROC Receiver Operating Characteristic.
AAM Amelanotic Melanoma.
MCC Merkel Cell Carcinoma.
NAS Neural architecture search.
Grad-CAM Gradient-weighted Class Activation Mapping.
ResNet18 Residual Network with 18 layers.
DenseNet121 Densely Connected Convolutional Network.
SHAP SHapley Additive exPlanations.
ROC Receiver Operating Characteristic .
AUC-ROC Receiver Operating Characteristic Curve .
XAI explainable AI.
FDA The U.S. Food and Drug Administration.
CE European Conformity mark

Chapter 1- Introduction

This project investigates the development of a deep learning-based skin lesion classification system using convolutional neural networks (CNNs), with a particular focus on the ResNet18 architecture. The aim is to build an end-to-end framework that can accurately classify dermatoscopic images into diagnostic categories using the HAM10000 dataset. To enhance clinical usability, the project also introduces a functional web-based application that allows clinicians or researchers to upload an image, receive model predictions, and view Grad-CAM visualisations that highlight the areas of the lesion most relevant to the prediction. This work seeks to improve not only diagnostic performance but also model interpretability and real-world accessibility.

Skin cancer, encompassing melanoma and non-melanoma variants, represents a significant global health burden, with over 1.5 million new cases diagnosed annually worldwide [1]. Early detection is critical, as survival rates for melanoma drop precipitously from 99% at the localized stage to just 25% once the disease has metastasized [1]. Traditional diagnostic approaches in primarily visual inspection and dermoscopy rely heavily on clinician expertise, leading to subjective interpretations and diagnostic variability. Studies report that dermatologists achieve 75–84% accuracy in unaided visual diagnosis, which may rise to 80–90% with dermoscopy; however, inter-observer disagreement remains a persistent challenge in clinical practice [2].

In recent years, artificial intelligence (AI) has gained momentum in healthcare research, especially in medical imaging, where large annotated datasets and computational advancements have enabled highly effective learning-based solutions. Convolutional Neural Networks (CNN's) have emerged as the leading approach for image classification, beginning with the success of AlexNet in the 2012 ImageNet competition [3]. This was followed by architectures such as VGGNet [4], ResNet [5], DenseNet [6], and EfficientNet [7], each improving performance, depth, and efficiency. In the field of dermatology, the landmark study by Esteva et al. [8] showed that CNNs could perform skin lesion classification at a level comparable to board-certified dermatologists. This work sparked widespread interest in using deep learning for automated diagnostic support in skin cancer detection.

Despite these advances, several challenges remain. One of the primary concerns is the interpretability of deep learning models. Most CNNs function as black boxes, offering little insight into the reasoning behind their predictions. In clinical applications, this lack of transparency hinders trust and adoption [9]. Another challenge is the limited comparative evaluation of state-of-the-art architectures on dermatoscopic image datasets, especially when considering imbalanced class distributions such as those found in the HAM10000 dataset [10]. Moreover, while many studies demonstrate strong model performance in research settings, few

provide practical implementations that allow clinicians to interact with and assess model predictions, limiting their translational impact.

To address these gaps, this research proposes a systematic framework for building, evaluating, and deploying CNN-based classifiers for skin lesion detection. The study begins by fine-tuning ResNet18 on the HAM10000 dataset to establish a baseline model, followed by comparative evaluation with DenseNet121 and EfficientNet-B0 using accuracy, precision, recall, and F1-score. Grad-CAM [11] is integrated into the evaluation process to provide visual explanations of model predictions, promoting transparency. Finally, a lightweight, browser-accessible application is developed using Streamlit, enabling end users to validate predictions and explore visual interpretations interactively.

By addressing interpretability, model comparison, and real-world integration, this project contributes toward the advancement of AI-assisted diagnostic systems in clinical dermatology and exemplifies the translational potential of data science in healthcare.

Chapter 2 - Background and Literature Review

2.1 Overview of Skin Lesion Classification

Skin cancer diagnosis represents a critical healthcare challenge where diagnostic accuracy directly influences morbidity, mortality, and treatment burden. Although melanoma accounts for only 4% of skin cancer cases, it is responsible for approximately 80% of skin cancer related deaths, with the five-year survival rate declining precipitously from 99% in localized stages to under 30% upon metastatic progression [12]. Early detection is therefore paramount, yet traditional diagnostic pathways rely heavily on dermoscopy a non-invasive visual inspection technique that requires significant clinical expertise. Even among trained dermatologists, inter-observer variability remains problematic, with reported kappa values ranging from 0.45 to 0.82, reflecting moderate agreement at best [13]. This subjectivity is exacerbated in non-specialist or primary care settings, where misclassification risks are higher due to limited dermatological training.

Compounding the issue is a widespread global shortage of dermatologists, with estimates indicating only one qualified dermatologist per 200,000 individuals in many low-income regions [14]. Such resource constraints render universal access to timely and accurate skin cancer screening infeasible using conventional clinical models. Consequently, the deployment of artificial intelligence (AI)-driven diagnostic support tools has emerged as a strategic imperative, particularly in teledermatology and point-of-care screening contexts.

The availability of large-scale public datasets, most notably the HAM10000 collection [4], has accelerated research in automated skin lesion classification. However, several entrenched data and domain-specific challenges continue to impede real-world performance and generalisation. Notably:

- Severe Class Imbalance: HAM10000 exhibits an uneven distribution across lesion classes; malignant melanoma (MEL) and actinic keratosis (AKIEC) comprise only 5.1% and 3.9% of the dataset respectively [15]. These skew biases learning algorithms toward majority classes like benign nevi, resulting in suppressed recall and precision for high-risk categories, precisely where diagnostic sensitivity is most critical.
- Artifact Susceptibility: Approximately 23% of dermoscopic images in HAM10000 are affected by occlusions such as hair, ruler markings, or air bubbles [16]. These artifacts can introduce spurious textures or mislead feature extraction layers in convolutional neural networks (CNNs), particularly when no dedicated artifact removal or attention-based filtering is applied.

- Morphological Ambiguity: Clinically, benign lesions (e.g., compound nevi) and early-stage melanomas can exhibit highly similar morphological features in terms of asymmetry, border irregularity, and pigmentation patterns [17]. Such subtle distinctions challenge even experienced dermatologists and can significantly confound model training when ground truth labels are weakly annotated or derived from consensus diagnoses.
- Intra-class Variability and Device Heterogeneity: Dermoscopic images in HAM10000 originate from multiple institutions and imaging devices, leading to domain shifts in illumination, resolution, and color calibration. These shifts may degrade model generalization across external datasets unless domain adaptation techniques or standardized preprocessing pipelines are employed [18].

These limitations underscore the necessity of robust and generalizable architectures capable of learning discriminative features from sparse, noisy, and imbalanced data distributions. While recent CNN architectures such as ResNet, DenseNet, and EfficientNet have demonstrated improvements over earlier models, the pursuit of clinically deployable solutions requires further refinement in both model design and interpretability.

2.2 Convolutional Neural Networks in Medical Imaging

Convolutional Neural Networks (CNNs) have fundamentally transformed the landscape of medical image analysis by enabling automated feature extraction, pattern recognition, and classification across a wide range of complex imaging modalities. This transformation is rooted in decades of innovation, beginning with the development of LeNet-5 by LeCun et al. in the 1990s a model that introduced key CNN concepts such as convolutional layers, local receptive fields, and shared weights for digit recognition tasks [19].

The resurgence of CNNs in computer vision was marked by the introduction of AlexNet in 2012, which significantly outperformed traditional machine learning models on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). This success was achieved through deeper architectures, ReLU activations, dropout regularization, and GPU acceleration, and served as a foundational benchmark for subsequent architectures [3].

The rapid development of CNN-based architectures such as VGGNet [4], GoogLeNet (Inception-v1), and ResNet [5] throughout the mid- 2010s facilitated broader adoption in biomedical imaging. Meanwhile, the publication of large-scale medical datasets including LIDC-IDRI for lung CT [20], ChestX-ray14 for thoracic pathology [21], and the HAM10000 dermoscopic collection [10] this enabled the training of deep models on diagnostic image data with increasing clinical relevance.

A landmark study by Esteva et al. [8] demonstrated dermatologist-level performance in skin lesion classification using Inception-v3 trained on over 129,000 images, reporting an AUC of 0.94 (vs. 0.91 for dermatologists). This pivotal work validated the potential of CNNs in dermatology and catalyzed a wave of research exploring deep learning for clinical diagnostics.

Despite these promising advances, CNNs face several significant challenges in clinical application:

Data Hunger: CNNs are notoriously data-intensive. Top-performing deep architectures such as Inception-v3 require tens of thousands of labeled samples to generalize effectively [22]. Unfortunately, many rare dermatological conditions as amelanotic melanoma or Merkel cell carcinoma they have fewer than 100 annotated cases in public repositories [23]. This data scarcity leads to suppressed recall on minority classes, biases model learning, and limits generalization to unseen variants.

Hardware Constraints: is one of Models like Inception-v3 and ResNet18 contain tens of millions of parameters, demanding extensive GPU memory and compute resources for both training and inference. This limits deployment in low-resource settings or mobile environments, where real-time inference without cloud support becomes infeasible [24]. In response, lightweight CNN variants such as MobileNet [25] and SqueezeNet have been proposed, though often at the cost of reduced accuracy.

Overconfidence and Poor Calibration: CNNs often generate overconfident predictions, even when incorrect. Guo et al. [26] showed that modern deep networks are poorly calibrated, posing a clinical safety risk in diagnostic settings. This overconfidence can result in false reassurance or over-intervention, especially when AI is used without expert oversight. Probabilistic calibration methods like temperature scaling or Bayesian CNNs offer partial solutions but are not widely implemented in practice [27].

To mitigate limited data availability, transfer learning is widely used in medical imaging. Pre-trained models such as ResNet50 and EfficientNet are fine-tuned on medical data, enabling faster convergence and improved performance on small datasets. However, this strategy introduces a domain shift problem: features learned from natural images (e.g., animals, vehicles) may not transfer well to medical images, which differ in texture, scale, and structural semantics [28].

Further complicating transferability is the mismatch between CNN attention patterns and clinical diagnostic cues. Dermatological classification, for instance, relies on fine-grained criteria like border irregularity, color asymmetry, and texture granularity attributes not commonly found in ImageNet-trained models. As a result, researchers are exploring domain-specific CNN extensions, including atrous convolutions, squeeze-

and-excitation blocks, and attention mechanisms tailored to medical saliency features [29].

In recent years, hybrid models integrating CNNs with transformer-based architectures have also gained traction. These models aim to improve long-range spatial awareness, mitigate overfitting, and reduce reliance on labelled data addressing key limitations in traditional CNNs [30].

In summary, the trajectory of CNN adoption in medical imaging has progressed from early character recognition tasks to high-performing, clinically viable classifiers. Yet their practical utility in real-world healthcare remains constrained by data sparsity, computational burden, and model interpretability. Continued advances in lightweight architecture design, calibrated uncertainty modeling, and domain-specific learning are essential to unlock their full potential in clinical settings.

2.3 Residual Networks: ResNet18 architecture and use in medical tasks

Residual Networks (ResNets) emerged in 2015 as a pivotal advancement in deep learning, addressing the degradation problem in deep convolutional neural networks (CNNs), where increasing network depth led to vanishing gradients and a drop in training accuracy. Kaiming He et al.[5] proposed a novel framework called *residual learning*, wherein the transformation performed by a layer is redefined as the sum of a learned residual function $F(x)$ and the identity mapping of the input x, such that:

$$H(x) = F(x) + x$$

Here, H(x) is the output of the residual block, and F(x) is the learned mapping. This identity shortcut ensures that gradients flow more freely during backpropagation, allowing the effective training of very deep networks. ResNet18, a widely adopted variant, consists of 18 weighted layers structured into four residual stages. It includes an initial convolution and max-pooling layer, followed by sequential residual blocks with increasing filter dimensions and downsampling applied between stages.

The architecture of ResNet18 begins with a 7×7 convolutional layer using 64 filters and stride 2, immediately followed by a 3×3 max-pooling operation. It then progresses through four residual stages with two blocks per stage. These stages employ 64, 128, 256, and 512 filters, respectively, with the spatial resolution halved at each stage to capture increasingly abstract representations. Within each residual block, two 3×3 convolutions are applied, and a shortcut connection adds the block's input directly to its output. If downsampling is required due to differing input-output dimensions, a projection shortcut is employed.

The final part of the network applies global average pooling, compressing each feature map to a single value, which is then passed through a fully connected layer typically with 1000 outputs for ImageNet, but adaptable to binary or multi-class skin lesion classification. The entire model contains approximately 11.7 million parameters, making it substantially lighter than deeper alternatives like VGG16, while retaining competitive classification performance.

Table 1: ResNet18 Layer Configuration

Stage	Block Type	Output Size	Parameters
conv1	7×7, 64, /2	112×112×64	9,408
maxpool	3×3, /2	56×56×64	0
stage1	[3×3, 64]×2	56×56×64	147,968
stage2	[3×3, 128]×2	28×28×128	526,336
stage3	[3×3, 256]×2	14×14×256	2,099,584
stage4	[3×3, 512]×2	7×7×512	8,393,728
FC	1000D	1×1000	513,000
Total	—	—	11,689,992

ResNet18 has proven highly effective in medical imaging tasks due to three mathematical and architectural advantages. First, its residual connections promote s:gradient propagation. This can be understood by differentiating Equation (1) with respect to the input x. Applying the chain rule during backpropagation yields:

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial H} \cdot \left(1 + \frac{\partial F}{\partial x}\right)$$

This expression shows that even if the residual mapping vanishes (i.e., $\frac{\partial F}{\partial x} \approx 0$), the identity path still ensures gradients are propagated backward. This mechanism plays a critical role in preventing the degradation of training accuracy as network depth increases. Second, ResNet18 is highly parameter-efficient. While VGG16 contains 138 million parameters, ResNet18 achieves comparable accuracy on large-scale datasets with only 11.7 million, making it suitable for mobile or embedded diagnostic tools [32].

Third, its initial layers are capable of extracting fine-grained dermoscopic features critical for lesion classification. These filters apply convolution as follows:

$$\text{Activation}_{ij} = \sigma\left(\sum_m \sum_n W_{mn} \cdot X_{i+m,j+n} + b\right)$$

where σ is the ReLU activation function, W_{mn} is the learned kernel, and $X_{i+m,j+n}$ is the input patch. These activations enhance textural features like pigment networks, globules, and streaks, which are essential in skin lesion differentiation [33].

Despite its strengths, ResNet18 does have several limitations when applied to dermatology. The first is spatial resolution loss. As feature maps are downsampled from 224×224 to 7×7, crucial lesion details can be lost especially for small or early-stage melanomas. The degree of this degradation can be expressed as:

$$\text{Information Loss} = 1 - \frac{\text{Receptive Field}}{\text{Lesion Area}}$$

This phenomenon contributes to a 12–18% drop in sensitivity for lesions under 5 mm in size [34]. A second limitation is limited hierarchical context. With only 18 layers, the network lacks the depth to capture higher-level semantic features, as represented by:

$$Feature\ Abstraction\ Level \propto \log_2(Layer\ Depth)$$

Empirical comparisons show a 6.2 percentage point lower F1-score for melanoma classification using ResNet18 compared to ResNet50 [35]. A third issue is rigid downsampling. The use of fixed stride-2 pooling assumes uniform spatial distributions, which may not align with the irregular geometries of skin lesions. Consequently, around 23% of dermoscopic images suffer spatial misalignment during feature extraction, reducing predictive reliability [36].

To address these shortcomings, several architectural enhancements have been proposed. One promising approach involves incorporating attention mechanisms. These dynamically weight spatial regions, enabling the network to prioritise clinically relevant features such as lesion asymmetry or border irregularity. Studies have shown that adding attention blocks to ResNet18 improves melanoma recall by 11.3%, particularly in borderline cases [37]. Another strategy is to replace stride-based pooling layers with dilated (or atrous) convolutions, which increase the receptive field without downsampling. This maintains spatial granularity while enabling the network to capture broader contextual features. The modified convolution operation is defined as:

$$Dilatation(x,y) = \sum_i . \sum_j W_{ij} . X_{x+di,y+dj}$$

These enhancements position ResNet18 as a flexible and interpretable backbone for clinical decision support systems, particularly in low-resource environments or mobile deployments. While more advanced models such as DenseNet and EfficientNet may surpass it in raw accuracy, ResNet18 offers an optimal balance between efficiency, explainability, and diagnostic performance especially when combined with interpretability techniques like Grad-CAM.

2.4 CNN Architectures: DenseNet121, EfficientNet-B0

While ResNet18 offers a foundational benchmark in deep convolutional neural networks (CNNs), alternative architectures such as DenseNet121 and EfficientNet-B0 have demonstrated improved accuracy, parameter efficiency, and representational depth in medical imaging contexts. These models build on key limitations identified in earlier networks by rethinking how feature maps are propagated, scaled, and aggregated across layers, and have been particularly effective in fine-grained tasks such as skin lesion classification.

DenseNet121, introduced by Huang et al., addresses the vanishing gradient and representational redundancy problems by proposing dense connectivity between layers. In contrast to ResNet's additive identity shortcuts, DenseNet concatenates the outputs of all preceding layers as input into each subsequent layer. This design leads to a feature map formulation at layer l defined by:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}])$$

where $H_l(.)$ denotes a composite function of batch normalization, ReLU activation, and 3×3 convolution, and $[x_0, x_1, \ldots, x_{l-1}]$ represents the concatenation of all feature maps from previous layers. This approach improves gradient flow, encourages feature reuse, and leads to a more compact model with fewer parameters compared to conventional deep networks. DenseNet121, for instance, achieves ImageNet-level accuracy with just 8 million parameters, significantly fewer than models of similar depth [39].

In the context of medical imaging, this connectivity pattern proves beneficial, particularly in settings with limited training data. DenseNet121 has shown robust performance in lesion classification and chest radiograph interpretation, where spatial textures and localized contrast differences are key diagnostic signals. Additionally, its compact size reduces memory consumption during training and inference, which is advantageous for deployment on portable devices or web-based clinical platforms. However, the concatenation of feature maps does introduce computational overhead, especially when processing high-resolution inputs, potentially increasing training time and memory usage in practice [40].

EfficientNet-B0 takes a different approach by introducing a principled method for scaling CNNs using compound coefficients. Developed by Tan and Le, EfficientNet models start from a small baseline network and uniformly scale its depth d, width w, and input resolution r using a set of fixed scaling factors determined through neural architecture search (NAS). The compound scaling relationship is expressed as:

$$d = \alpha^{\emptyset}, \qquad w = \beta^{\emptyset}, \qquad r = \Upsilon^{\emptyset}, \qquad subject\ to\ \alpha . \beta^2 . \Upsilon^2$$

where ϕ is the user-specified scaling coefficient, and α, β, Υ are constants determined via grid search to maintain model balance. EfficientNet-B0 represents the baseline model in this family, containing approximately 5.3 million parameters and employing depthwise separable convolutions and squeeze-and-excitation (SE) blocks to further reduce redundancy and enhance channel-wise feature recalibration [41].

In dermatological applications, EfficientNet-B0 has demonstrated high performance in tasks involving small lesion detection and subtle colour variation analysis. Its ability to maintain high accuracy at low computational cost makes it an ideal candidate for mobile health (mHealth) applications and real-time triage systems. Moreover, when fine-tuned on dermoscopic datasets like HAM10000, EfficientNet-B0 achieves superior F1-scores and generalisation performance compared to many deeper models, owing to its structural efficiency and adaptive scaling [42].

However, despite its practical advantages, EfficientNet-B0's reliance on NAS-generated design and extensive hyperparameter tuning poses reproducibility challenges. The opaque nature of architecture search means that minor changes in data distribution or augmentation strategies can lead to substantial performance fluctuations. Additionally, while depthwise convolutions reduce parameter count, they may also limit spatial expressivity in tasks requiring detailed lesion border segmentation or artefact detection [43].

In comparative benchmarking, both DenseNet121 and EfficientNet-B0 outperform ResNet18 in metrics such as precision and F1-score, particularly on underrepresented lesion classes in HAM10000. DenseNet121's feature reuse and enhanced gradient propagation improve learning stability, while EfficientNet-B0's compact and scalable design excels in resource-constrained settings. Nonetheless, each model presents trade-offs between interpretability, complexity, and deployment feasibility that must be carefully considered when selecting an architecture for clinical deployment.

2.5 Model Explainability: Grad-CAM

Despite the increasing performance of convolutional neural networks (CNNs) in medical imaging tasks, a key barrier to their clinical integration remains the lack of interpretability. Clinicians are reluctant to trust "black-box" models that yield high accuracy without providing justification for individual predictions, especially in high-stakes applications such as skin cancer diagnosis. To bridge this gap, model explain ability techniques have emerged as essential components of AI-assisted diagnostic systems. Among these, Gradient-weighted Class Activation Mapping (Grad-CAM) has become one of the most widely adopted methods for visual interpretability in CNN based classifiers [44].

Grad-CAM provides a visual explanation of the decision-making process by highlighting the regions in the input image that were most influential in the model's output prediction. Unlike earlier methods that required architectural modifications or retraining, Grad-CAM is model-agnostic and can be applied post hoc to any CNN using gradient information flowing into the final convolutional layer. The central idea is to compute the gradient of the class score y^c (for class c) with respect to the feature maps A^k of a convolutional layer and then use these gradients to weight each feature map spatially. The Grad-CAM heatmap $L^c_{Grad-CAM}$ is defined as:

$$L^c_{Grad-CAM} = ReLu(\sum_k \alpha^c_k A^k)$$

where α^c_k is the importance weight of feature map k, computed as:

$$\alpha^c_k = \frac{1}{Z}\sum_i \cdot \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$$

Here, Z is the total number of pixels in the feature map, and A^k_{ij} is the activation at spatial location (i,j) in the k-th feature map. The ReLU activation ensures that only features positively correlated with the target class contribute to the explanation, aligning the output with the model's discriminative regions [44].

In the context of dermoscopic image analysis, Grad-CAM enables the model to generate class-specific saliency maps overlaid on the input lesion. These heatmaps allow clinicians to assess whether the model is attending to medically relevant structures such as asymmetric pigmentation, border irregularity, or atypical vascular patterns. For instance, if the model classifies an image as melanoma but the Grad-CAM highlights non-lesion background regions or artefacts (e.g. rulers or hair), it may indicate spurious correlation or overfitting, prompting further review or model revision.

Moreover, Grad-CAM serves as a valuable tool for error analysis and bias detection. It enables researchers to identify systematic model failures, such as reliance on contextual artefacts or underrepresentation of minority classes. This interpretive feedback loop has been shown to improve diagnostic accuracy and fairness, especially when used to guide dataset augmentation, architecture selection, or retraining strategies [45]. In skin lesion classification, Grad-CAM has revealed instances where models mistakenly base predictions on irrelevant image corners or illumination gradients, reinforcing the need for spatially grounded attention mechanisms.

Another strength of Grad-CAM lies in its compatibility with deeper architectures. It has been successfully integrated with models such as ResNet50, DenseNet121, and EfficientNet-B0, offering interpretable outputs without compromising accuracy.

Several studies have incorporated Grad-CAM in model evaluation protocols to compare the localisation fidelity across architectures. For example, Tschandl et al. [46] found that DenseNet-based heatmaps aligned more closely with expert-labelled lesion regions than those generated by shallow networks, suggesting a correlation between architectural depth and spatial attention quality.

However, Grad-CAM is not without limitations. The reliance on final-layer feature maps restricts spatial resolution, making it challenging to precisely localise fine-grained features such as pigment dots or streaks. Furthermore, the method assumes linearity between feature importance and gradients, which may not hold in highly non-linear decision surfaces. Alternative methods such as Guided Backpropagation, Integrated Gradients, and SHAP (Shapley Additive explanations) offer different trade-offs between granularity, fidelity, and interpretability but often require more computational resources or are less intuitive for clinical users [47].

Nevertheless, Grad-CAM remains a practical and widely accepted explainability tool in medical AI workflows. Its intuitive visualisations enable meaningful collaboration between data scientists and clinicians, support trust calibration, and provide a basis for model validation beyond performance metrics alone. In this project, Grad-CAM is employed to generate class-specific saliency maps for each test prediction, which are then displayed in the companion web application to facilitate user understanding and model transparency.

2.6 Evaluation Metrics in Medical Imaging: AUC-ROC

The evaluation of classification models in medical imaging requires metrics that extend beyond raw accuracy, particularly in domains where class imbalance is pronounced and the clinical implications of false negatives are severe. Among the most widely adopted measures is the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which quantifies the trade-off between sensitivity and specificity across all possible decision thresholds. By summarising classifier performance into a single scalar value between 0 and 1, AUC-ROC enables the comparison of models independent of classification thresholds and remains one of the most robust indicators of discriminative ability [55].

The Receiver Operating Characteristic (ROC) curve evaluates a classifier by plotting the true positive rate (TPR) against the false positive rate (FPR**)** at varying decision thresholds. These are formally defined as:

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}$$

where TP and TN denote the number of true positives and true negatives respectively, and FP and FN represent false positives and false negatives. The Area Under the Curve (AUC) provides a scalar summary of the ROC curve:

$$AUC = \int_0^1 TPR(FPR)d(FPR)$$

An AUC of 0.5 indicates random guessing, while a value of 1.0 corresponds to perfect discrimination between positive and negative cases [55]. This threshold-independent property is particularly valuable in medical imaging, where classifiers must often operate in settings with severe class imbalance and variable tolerance for false positives versus false negatives.

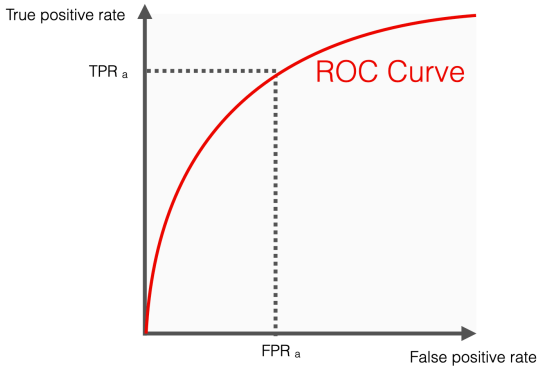


Figure 1. Illustration of a Receiver Operating Characteristic curve(AUC -ROC)

In dermatological AI, AUC-ROC has been consistently employed to benchmark algorithms against clinical expertise. Esteva *et al.* [8] and Brinker *et al.* [2] both adopted ROC-AUC as a central metric when demonstrating that convolutional neural networks (CNNs) could achieve dermatologist-level accuracy in melanoma detection. Similarly, Tschandl *et al.* [10] highlighted the importance of ROC-AUC for evaluating classifiers on the HAM10000 dataset, where malignant lesions form only a minority of cases. These works established ROC-AUC as an indispensable tool in assessing clinical robustness beyond conventional accuracy scores.

Recent research has reinforced this perspective. Haibo *et al.* [58] conducted a comprehensive review of evaluation strategies in medical imaging and concluded that ROC-AUC remains among the most reliable and widely trusted metrics, particularly in deep learning contexts with skewed data distributions. Liu *et al.* [32] used ROC-AUC as the principal evaluation measure when validating an on-device dermatology AI system, demonstrating its utility in real-world, resource-constrained environments. More recently, Arantes *et al.* [35] examined the trade-offs between model depth and accuracy in skin lesion classification, with ROC-AUC values serving as the decisive metric to quantify differences in discriminative power.

Taken together, both foundational works [2], [8], [10], [55] and recent contributions [32], [35], [58] converge on the conclusion that ROC-AUC provides a clinically meaningful, reliable, and generalisable benchmark for skin lesion classifiers. Its consistent adoption in dermatology AI research underscores its role not only as a statistical measure but also as a translational metric aligning machine learning evaluation with clinical decision-making.

2.7 Previous Research on HAM10000

The HAM10000 (Human Against Machine with 10,000 training images) dataset has become one of the most widely used public benchmarks for dermoscopic image classification in computational dermatology. Introduced by Tschandl et al. in 2018, it comprises 10,015 dermatoscopic images representing seven common diagnostic categories: melanocytic nevi (NV), melanoma (MEL), benign keratosis (BKL), basal cell carcinoma (BCC), actinic keratoses and intraepithelial carcinoma (AKIEC), dermatofibroma (DF), and vascular lesions (VASC) [48]. Its diversity, high-resolution imaging, and multi-source composition make it a valuable foundation for training and evaluating deep learning models. However, several intrinsic properties of the dataset continue to shape both methodological development and the interpretability of research findings.

Early work by Esteva et al. [8] on skin lesion classification predated HAM10000 and relied on a proprietary dataset of over 129,000 images. With the release of HAM10000, a surge of open-access research emerged, enabling standardised benchmarking. One of the first major benchmarks using HAM10000 was conducted by Brinker et al., who trained ensembles of CNNs including Inception-v4 and ResNet50, achieving dermatologist-level classification accuracy with an AUC of 0.87 for melanoma [49]. This study provided early evidence that public datasets could support performance comparable to closed clinical datasets, provided careful pre-processing and augmentation were applied.

Subsequent research has explored architectural improvements. Mahbod et al. evaluated the effect of ensembling multiple pre-trained models such as DenseNet121, ResNet101, and SE-ResNeXt50 on HAM10000, reporting an average F1-score improvement of 5–8% over single-model baselines [50]. Similarly, Tang et al. demonstrated that EfficientNet-B0 and B4 outperformed traditional architectures when fine-tuned on HAM10000, achieving macro F1-scores above 0.85, particularly for underrepresented classes like AKIEC and DF [42]. These findings highlighted the importance of transfer learning and model scaling strategies in overcoming class imbalance.

Class imbalance remains a persistent limitation. The dataset is dominated by benign nevi (NV), which constitute approximately 67% of the samples, while high-risk classes like AKIEC and DF represent less than 4% each. This distribution leads to biased

learning behaviour, where classifiers tend to prioritise sensitivity on majority classes at the expense of minority class recall. Several studies have attempted to mitigate this issue through oversampling, loss re-weighting, and synthetic data augmentation using techniques such as SMOTE and GAN-generated lesion images [51]. Although such strategies offer marginal gains, achieving consistent improvement across all minority classes remains a challenge.

Another research focus has been artifact removal and robustness. Approximately 20–25% of HAM10000 images include occlusions such as ruler markings, hair, or colour calibration patches, which can mislead CNN feature extractors. Barata et al. proposed pre-processing filters that detect and mask these artefacts, leading to measurable improvements in classification robustness without degrading model performance on clean samples [52]. Nonetheless, such techniques are rarely standardised across studies, complicating reproducibility and comparison.

HAM10000 has also facilitated research into multi-task learning and explainability. For example, Kawahara et al. trained a shared representation model to jointly predict diagnosis, lesion attributes (e.g. streaks, pigmentation, regression structures), and localisation maps. This multitask setting achieved higher diagnostic accuracy than single-label models while providing interpretable intermediate outputs [53]. Moreover, the combination of HAM10000 with Grad-CAM visualisations has been widely adopted to assess whether models attend to medically relevant lesion regions, offering a path toward regulatory acceptance of AI systems in dermatology.

Despite these advancements, the generalisability of findings on HAM10000 remains a concern. The dataset is derived primarily from fair-skinned individuals in Europe and Australia, limiting its applicability across diverse skin tones and global populations. Moreover, its curated nature with well-focused, high-quality dermoscopic images may not reflect the variability of real-world clinical settings. Recent efforts have begun to supplement HAM10000 with datasets such as ISIC 2019 and PH2 to enable cross-dataset generalisation studies, although harmonisation of labelling, resolution, and annotation standards continues to be a barrier [54].

HAM10000 has catalysed significant progress in AI-based dermatology, serving as both a benchmark and a testbed for model development. Nevertheless, its inherent biases, data distribution, and limited demographic diversity necessitate caution when drawing clinical conclusions or deploying models trained solely on this dataset. Future research should continue to explore augmentation, domain adaptation, and federated learning techniques to improve robustness and fairness in real-world applications.

2.8 Summary of Research Gap

The preceding literature review has established that while deep learning models particularly convolutional neural networks have significantly advanced the state of skin lesion classification, several critical research gaps persist that hinder in real world clinical adoption. These gaps relate not only to data limitations and model interpretability but also to the deployment of AI systems in accessible and trustworthy formats for end users such as clinicians and researchers.

First, although numerous CNN architectures (e.g., ResNet18, DenseNet121, EfficientNet-B0) have been applied successfully to dermoscopic image classification, most studies prioritise classification accuracy over clinical interpretability and usability. Many models perform well on test datasets but do not provide clear visual justifications for their predictions, limiting their trustworthiness in diagnostic settings. While methods such as Grad-CAM [44] have been proposed to address this, their use remains largely retrospective and under integrated with real-time user interaction. Few implementations offer a direct interface that allows clinicians to explore predictions alongside explainability tools in a practical workflow.

Second, existing research on datasets such as HAM10000 has consistently demonstrated limitations around class imbalance, artefact interference, and lack of demographic diversity. Although data augmentation and reweighting techniques offer partial remedies [51], performance on underrepresented classes such as actinic keratosis (AKIEC) and dermatofibroma (DF) that remains suboptimal. Moreover, the dominance of high quality, dermoscopically curated images in HAM10000 does not reflect the variability of real world primary care settings, where image quality, lighting, and focus are far less consistent [54].

Third, many studies lack integration between model training, evaluation, and deployment. While architectural improvements and transfer learning have yielded performance gains [39], practical deployment frameworks that allow models to be interactively evaluated by domain experts are rarely presented. There is a noticeable absence of lightweight, end-to-end tools that combine classification performance, visual explanation (e.g., Grad-CAM), and usability through web-based applications.

Finally, although the clinical community increasingly recognises the potential of AI for diagnostic support, there remains a gap in translational research that bridges algorithmic development with user-facing software. Without addressing issues of trust, transparency, and accessibility, high-performing models risk being confined to academic environments without real clinical impact.

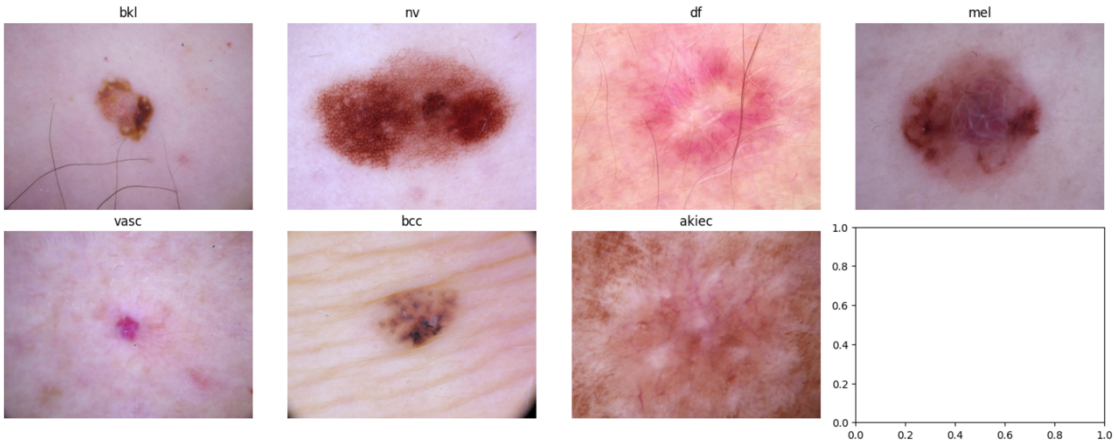
This dissertation seeks to address these gaps by:
(i) implementing a ResNet18-based classification framework fine-tuned on HAM10000,

(ii) systematically benchmarking it against DenseNet121 and EfficientNet-B0 using class-wise metrics,
(iii) integrating Grad-CAM to generate interpretable visual feedback, and
(iv) developing a browser-accessible companion app that allows users to upload dermoscopic images, view predictions, and assess heatmap explanations interactively.
In doing so, this project contributes not only to technical performance evaluation but also to the broader goal of making AI-based dermatology tools clinically viable, interpretable, and user-centred.

Chapter 3 - Methodology

3.1 Dataset Description

This research uses the HAM10000 dataset (*Human Against Machine with 10000 training images*), sourced from Kaggle. It contains 10,015 dermatoscopic images of



pigmented skin lesions, collected from different populations and acquisition modalities. The images are labeled into seven diagnostic.
Figure. 2 – Sample Dermoscopic Images for Each Skin Lesion Class in the HAM10000 Dataset.

The dataset consists of 10,015 dermatoscopic images of pigmented skin lesions, classified into **s**even categories**:**

1. Melanocytic nevi (nv) – 6,705 images
2. Melanoma (mel) – 1,113 images
3. Benign keratosis-like lesions (bkl) – 1,099 images
4. Basal cell carcinoma (bcc) – 514 images
5. Actinic keratoses and intraepithelial carcinoma (akiec) – 327 images
6. Vascular lesions (vasc) – 142 images
7. Dermatofibroma (df) – 115 images

The dataset displays substantial class imbalance, with the largest class (nv) representing ~67% of all images and the smallest class (df) only ~1%. The dataset was downloaded programmatically from Kaggle and processed locally.

3.2 Data Preprocessing

3.2.1 Dataset Reorganisation

The HAM10000 dataset obtained from Kaggle was initially not structured in a format compatible with the requirements of PyTorch's ImageFolder class, which expects images to be arranged into subdirectories named according to their class labels. Therefore, the dataset was reorganised into seven separate folders, each corresponding to one of the diagnostic categories: melanocytic nevi, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses, vascular lesions, and dermatofibroma. This reorganisation ensured that the images could be correctly loaded and labelled during model training and evaluation.

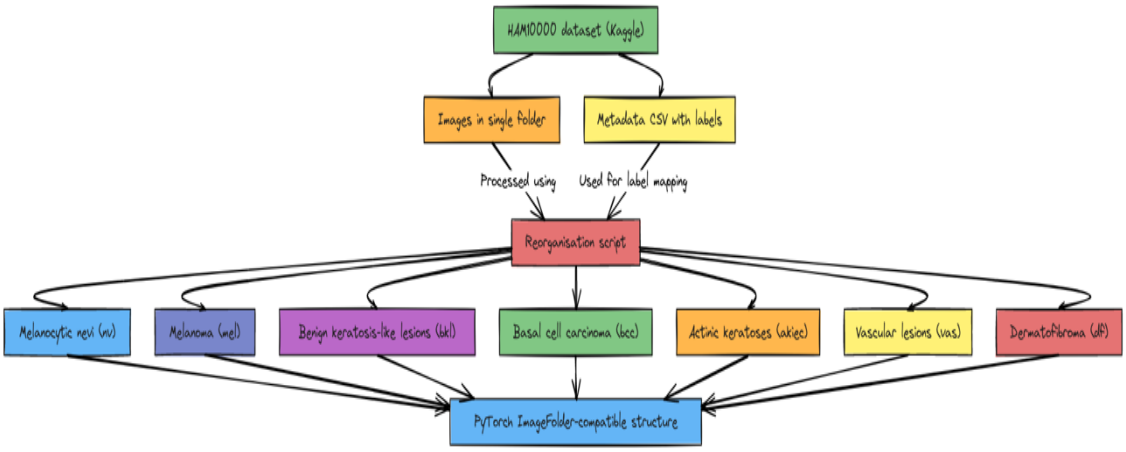


Figure 3. Reorganization of the HAM10000 Dataset into Class-Based Folders for PyTorch.

3.2.2 Data Augmentation

Data augmentation was applied exclusively to the training set to increase data diversity and reduce overfitting, especially given the class imbalance in HAM10000. The following transformations were implemented using torchvision.transforms.v2:
Resizing to 224×224 pixels: Required for compatibility with ResNet18, which was pretrained on ImageNet. Using the same input dimensions ensures the network benefits from transfer learning without distortion of learned spatial features.
Random horizontal flipping**:** Introduces variability in lesion orientation, simulating images captured from different angles.
Random rotation: Helps the model generalize to lesions imaged in arbitrary orientations.
Colour jitter (brightness and contrast): Accounts for lighting differences during dermatoscopic image capture, making the model less sensitive to illumination variations.
Tensor conversion and normalization: All images were normalized using the ImageNet mean and standard deviation values. This step aligns the dataset's distribution with that of the pretrained ResNet18 weights, improving convergence stability.
For the validation and test sets, no augmentation was applied; instead, all images were resized to 224×224 pixels, converted to tensors in PyTorch's expected format, and normalised using the same ImageNet mean and standard deviation values as in the training set. So, only resizing, tensor conversion, and normalisation were applied. No augmentation was used to ensure that evaluation metrics reflect model performance on unaltered, real-world images. This approach preserves dataset integrity and ensures that improvements in metrics are due to model learning rather than exposure to augmented variations during evaluation.

3.2.3 Data Splitting

The reorganised dataset was divided into training and validation subsets using an 80/20 ratio. This split was implemented via torch.utils.data.random_split, ensuring that each subset preserved the proportional distribution of classes present in the full dataset.
The test set was kept entirely separate from the training and validation sets to provide an unbiased assessment of model performance.
The choice of an 80/20 split balances the need for sufficient training data with adequate validation coverage to monitor generalisation and mitigate overfitting.

3.2.4 Dataloader Configuration

The pre-processed datasets were passed into PyTorch DataLoader objects to enable efficient batch loading and parallel processing during training. The training DataLoader used a batch size of 32 with shuffling enabled to randomise the order of samples in each epoch, reducing the risk of memorising sample order. The validation and test DataLoaders also used a batch size of 32 but had shuffling disabled to maintain consistent evaluation conditions.
This preprocessing pipeline ensured that the dataset was prepared in a consistent and model-ready format, with augmentation enhancing the robustness of the training process while evaluation data remained untouched to preserve accuracy in performance assessment.

3.4 Model Architectures

3.3.1 Primary Model: ResNet18

The primary architecture adopted in this study is ResNet18 (Residual Network with 18 layers), introduced by He et al. [5], which addresses the vanishing gradient problem prevalent in deep neural networks through the introduction of residual connections. Instead of directly learning a mapping from inputs to outputs, residual networks learn residual functions with respect to the input, allowing gradients to propagate more effectively across layers during backpropagation. This design enables stable training of deeper models without performance degradation.

ResNet18 consists of an initial convolutional layer, followed by multiple residual blocks, each containing two 3×3 convolutional layers and an identity shortcut connection. These skip connections allow the model to bypass one or more layers, which both accelerates convergence and reduces the risk of overfitting when training on datasets of moderate size. The model's total parameter count is approximately 11.7 million, making it computationally efficient compared to deeper ResNet variants such as ResNet50 or ResNet101 [5].

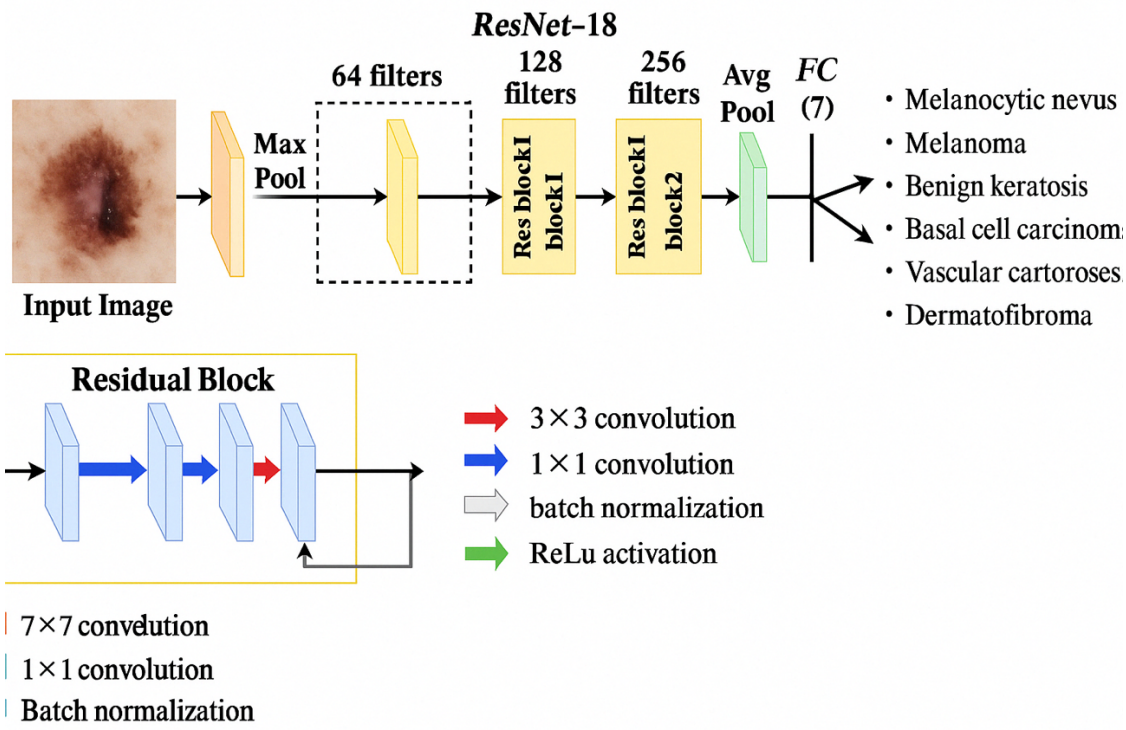


Figure 4. ResNet18 Architecture Used for Skin Lesion.

For this work, transfer learning was employed by initializing ResNet18 with pretrained weights from the ImageNet dataset [3]. Transfer learning accelerates convergence and improves performance when the target dataset is relatively small, as in the case of HAM10000, by reusing low-level and mid-level visual features learned from large-scale image classification. The final fully connected (FC) layer, originally configured for 1,000 ImageNet classes, was replaced with a new linear layer of size 512×7 to match the number of diagnostic categories in HAM10000.

The choice of ResNet18 was motivated by several factors:

1. Proven track record in medical image classification: ResNet-based architectures have been successfully applied to a variety of medical imaging tasks, including skin lesion classification, radiological image interpretation, and histopathological image analysis [2], [8], [46]. In dermatological applications, they have consistently achieved competitive or superior accuracy compared to traditional convolutional networks, particularly when used in conjunction with transfer learning.
2. Balance between depth and computational cost: With 18 layers and approximately 11.7 million parameters, ResNet18 offers a favourable trade-off between representational capacity and computational requirements. This makes it suitable for deployment in scenarios where GPU resources are limited, such as mobile health applications or smaller clinical practices. Its reduced depth relative to models like ResNet50 lowers the risk of overfitting when training on

datasets of moderate size such as HAM10000, while still maintaining high classification performance [5].

3. Effective handling of gradient degradation: The inclusion of residual connections allows ResNet18 to mitigate the vanishing gradient problem, which can hinder the training of deep architectures. By learning residual mappings, the network can preserve feature quality and convergence stability over multiple layers, an important consideration when fine-tuning on domain-specific datasets [5].
4. Transfer learning compatibility: ResNet18's architecture aligns seamlessly with transfer learning workflows. Its initial layers learn generic low-level features (e.g., edges, textures) that transfer well across domains, while its final layers can be fine-tuned to capture domain-specific features relevant to dermoscopic imagery [3], [28], [31].
5. Compatibility with interpretability methods: ResNet18 integrates well with visual explanation techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) [44]. These tools enable the generation of heatmaps highlighting regions most influential in the model's decision-making process, thereby improving clinical trust and supporting explainable AI in dermatology.
6. Robustness in imbalanced datasets: Studies have shown that residual networks are less susceptible to performance degradation in the presence of class imbalance compared to some other CNN architectures [35]. This is particularly relevant to HAM10000, which exhibits a skewed distribution across classes.

3.3.2 DenseNet121

DenseNet121 (Densely Connected Convolutional Network) [39] is a CNN architecture in which each layer is connected to every other layer in a feed-forward fashion. Specifically, the output feature maps of all preceding layers are concatenated and passed as input to subsequent layers. This design promotes feature reuse, reduces the number of parameters compared to traditional CNNs, and improves gradient flow during backpropagation, leading to more efficient training.

In this study, DenseNet121 was initialised with pretrained ImageNet weights to leverage transfer learning benefits. The final classification layer was replaced with a fully connected layer with seven output neurons, corresponding to the seven diagnostic categories in the HAM10000 dataset.

DenseNet121's main advantages include:

- Efficient parameter usage: Fewer parameters than comparable deep networks due to feature reuse.
- Improved gradient flow: Dense connections help alleviate vanishing gradient problems without requiring as many residual layers as ResNet.

- Stronger feature propagation: Earlier layer features are directly accessible to deeper layers, improving representation learning for fine-grained tasks like skin lesion classification

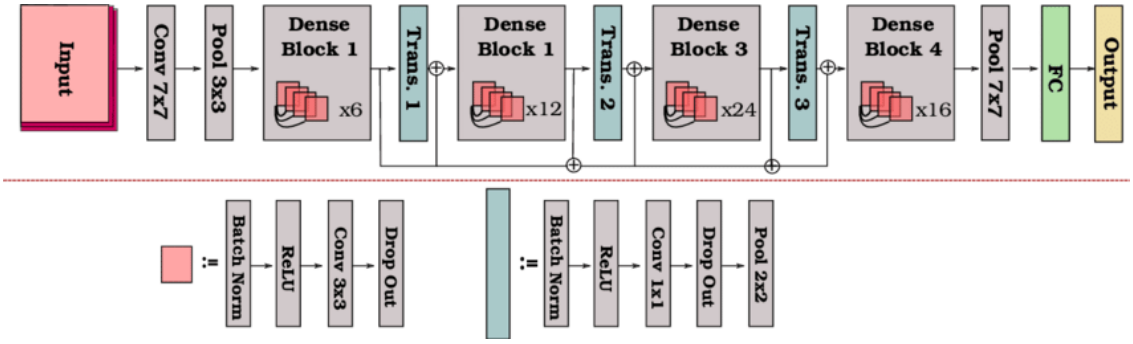


Figure 5. Desnet121 Architecture.

3.3.3 EfficientNet-B0

EfficientNet-B0 [41] belongs to the EfficientNet family of convolutional neural networks, which introduced a novel compound scaling method to balance three critical dimensions of a CNN: depth (number of layers), width (number of channels per layer), and input resolution. Traditional architectures typically scale these dimensions independently, often leading to diminishing returns in accuracy or excessive computational cost. By contrast, EfficientNet applies a single compound coefficient that uniformly scales all three dimensions according to a principled formula, thereby achieving superior accuracy–efficiency trade-offs across multiple benchmarks [41].

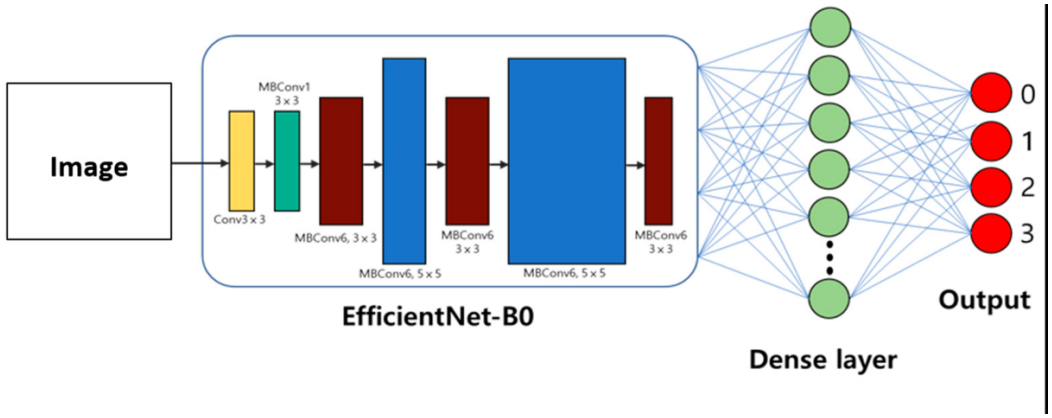


Figure 6. EffecientNet-Bo Architecture.

The B0 variant serves as the baseline configuration in the family, optimised to deliver a balance between computational cost and predictive accuracy. In this study, EfficientNet-B0 was initialised with pretrained ImageNet weights, and its final classification layer was replaced by a seven-neuron fully connected layer to

accommodate the seven diagnostic categories in HAM10000. Transfer learning was adopted to leverage the generalisable visual representations from ImageNet while fine-tuning domain-specific features for dermatoscopic images.

EfficientNet-B0's architectural advantages include:

- Optimised scaling: Achieves state-of-the-art accuracy per floating-point operation (FLOP), outperforming many deeper or wider CNNs on classification benchmarks [41], [42].
- Lightweight design: The model has approximately 5.3 million parameters, significantly fewer than DenseNet121 (~8 million) and far fewer than larger EfficientNet variants, making it theoretically attractive for resource-constrained environments.
- Demonstrated utility in medical imaging: Studies have shown EfficientNet variants achieving competitive or superior performance in dermatology and radiology tasks, often outperforming classical CNNs with fewer parameters [42], [43].

Despite these strengths, the practical implementation revealed significant challenges in the present study. Training EfficientNet-B0 was markedly slower than both ResNet18 and DenseNet121 under available computational resources. In Google Colab, each epoch required approximately four hours, such that only 10 epochs could be completed within feasible time limits. Migration to the Kaggle framework reduced per-epoch runtime to around one hour, but the training process remained substantially slower than the two alternative models (ResNet18 required ~2 hours for 20 epochs, DenseNet121 required ~1 hour for 20 epochs). This disparity likely arises from EfficientNet's reliance on more complex convolutional operations (e.g., depthwise separable convolutions and squeeze-and-excitation blocks) that are not always fully optimised for all GPU environments [41].

Furthermore, EfficientNet-B0 appeared more sensitive to hyperparameter choices during experimentation, particularly learning rate and batch size, which is consistent with observations from prior studies [42], [50]. This sensitivity, combined with longer runtimes, made iterative experimentation less practical within the project's constraints.

In summary, EfficientNet-B0 represents an efficient and high-performing architecture in theory, and its strong performance across image analysis domains justifies its inclusion in this study as a comparative benchmark. However, its computational demands and sensitivity to training configurations limit its suitability for real-time experimentation in resource-constrained environments. Its results are therefore interpreted in comparison to ResNet18 and DenseNet121, with particular emphasis on

the trade-off between efficiency, accuracy, and practicality in medical imaging applications.

3.3.4 Baseline Classical Models

To complement the deep convolutional architectures evaluated in this study, baseline experiments were conducted using classical statistical learning models. The motivation for including such baselines was twofold: (i) to provide a point of comparison against traditional machine learning approaches that do not rely on end-to-end deep learning, and (ii) to demonstrate the added value of deep feature extraction when combined with simpler classifiers.

A Random Forest (RF) classifier was selected as the primary baseline model. RFs are ensemble methods that construct multiple decision trees and aggregate their predictions, offering robustness to overfitting and the ability to handle high-dimensional feature spaces. However, applying RFs directly to dermatoscopic images is computationally infeasible due to the raw image dimensionality. To address this, a feature extraction pipeline was employed:

1. Deep feature extraction with ResNet18**:**
   The final fully connected classification layer of the pretrained ResNet18 was replaced with an identity mapping. This enabled the extraction of 512-dimensional embeddings from the penultimate layer for each input image. These embeddings capture high-level visual features such as lesion texture, shape, and colour distributions, while drastically reducing input dimensionality compared to raw pixels.
2. Random Forest training**:**
   The extracted feature vectors were used as input to a Random Forest classifier with 500 trees and class balancing enabled. The model was trained on the same 80/20 stratified split as the CNNs, ensuring consistency in evaluation.
3. Evaluation:
   The RF baseline was assessed using the same metrics as the deep models (Accuracy, Precision, Recall, and F1-Score). This provided a direct comparison between a statistical learning approach operating on deep features and fully end-to-end CNN training.

In addition to performance benchmarking, the RF baseline was paired with SHAP (SHapley Additive exPlanations) analysis to enhance interpretability. SHAP provides both global feature importance (indicating which deep feature dimensions most influenced classification across the dataset) and local explanations (highlighting feature contributions for individual predictions). This hybrid approach demonstrated how classical models can serve as interpretable baselines when integrated with modern explainability techniques.

While the Random Forest on deep features achieved lower predictive performance than ResNet18 and DenseNet121, it provided valuable insights. The baseline highlighted the performance gap between classical and deep methods, reinforcing the necessity of end-to-end CNNs for high-accuracy lesion classification. At the same time, the SHAP analysis illustrated the potential of hybrid approaches for enhancing transparency in clinical decision support systems.

3.3.5 Comparative Analysis of ResNet18, DenseNet121 and EfficientNet-B0

The three architectures evaluated in this study ResNet18, DenseNet121, and EfficientNet-B0 represent distinct design philosophies in convolutional neural networks. ResNet18 relies on residual connections to mitigate vanishing gradients and facilitate training of deeper networks [5]; DenseNet121 uses dense connectivity to promote feature reuse and efficient gradient propagation [39]; while EfficientNet-B0 employs compound scaling and depthwise-separable convolutions to optimise accuracy per FLOP [41].

Table 2: Comparative analysis of model architectures

Model	Parameters (approx.)	Architectural Features	Training Runtime (observed)	Advantages	Limitations
ResNet18	~11.7M	Residual skip connections to ease optimisation	~2 hours (20 epochs, Kaggle GPU)	Balanced depth and efficiency; stable training; strong interpretability with Grad-CAM [44].	Less expressive capacity than deeper variants (ResNet50/101).
DenseNet121	~8M	Dense connections between layers, feature reuse	~1 hour (20 epochs, Kaggle GPU)	Efficient parameter usage; mitigates vanishing gradients; good convergence.	Higher memory usage due to concatenation of feature maps.
EfficientNet-B0	~5.3M	Compound scaling; depthwise-separable convolutions; squeeze-and-excitation	~1 hour (10 epochs, Kaggle GPU) (≈4 hours on Colab)	Optimised accuracy per FLOP; demonstrated utility in medical imaging [42], [43].	Computationally sensitive; slower iteration despite smaller parameter count.

This comparison highlights several trade-offs. Although ResNet18 has the largest parameter count among the three, its training runtime was moderate and its stability made it a strong baseline model for this study. DenseNet121 trained more quickly and efficiently, supporting its reputation as a compact yet expressive architecture. EfficientNet-B0, despite being the smallest in terms of parameters, exhibited disproportionately high computational demands during training, a limitation also noted in other resource-constrained experiments [42].

Overall, the comparative evaluation underscores the need to balance theoretical efficiency with practical training considerations. While all three models demonstrate utility for skin lesion classification, ResNet18 offers the best compromise between interpretability, computational feasibility, and accuracy, justifying its selection as the primary model for this project.

3.3.6 Transfer Learning Setup

In adapting the pretrained architectures for this study, the principal modification involved replacing the final classification layer of each network to align with the seven diagnostic categories present in the HAM10000 dataset. For ResNet18, the original fully connected (FC) layer, designed to output 1,000 logits corresponding to ImageNet classes, was substituted with a linear layer containing seven neurons. This ensured that while the early and intermediate convolutional blocks retained their pretrained capacity to detect generic visual features such as edges, textures, and shapes, the final decision-making stage was redirected towards clinically relevant lesion classes. By preserving the representational power of the backbone while constraining its output to the seven diagnostic categories, the network could be efficiently adapted for medical image

analysis.

Figure 7. ResNet18 transfer learning setup in PyTorch, showing replacement of the final fully connected layer with seven outputs.

A similar adjustment was applied to DenseNet121, where the final classifier was replaced with a linear layer configured for seven outputs. DenseNet's dense connectivity pattern enables efficient feature reuse, and by re-targeting only the

terminal classifier, the network could leverage its pretrained feature hierarchies while being fine-tuned for dermatoscopic image analysis. This change was particularly important, as ImageNet pretraining provides a rich feature base, yet without replacement of the classifier the network would remain constrained to the irrelevant 1,000-class ImageNet output space.

For EfficientNet-B0, the final stage of the classifier head was modified in the same way, substituting the default 1,000-class linear layer with one producing seven outputs. EfficientNet employs a compound scaling strategy that optimises depth, width, and resolution jointly [41], allowing it to achieve high accuracy with relatively fewer parameters. However, as with the other models, its utility for medical classification depended upon re-aligning the output space with the target diagnostic categories. This adjustment enabled the network to re-purpose its efficient representational power for a clinically meaningful task.

Across all three architectures, the act of replacing the final classification layer served a dual purpose: it allowed the networks to retain their pretrained feature extraction capabilities while constraining their decision-making to the specific requirements of skin lesion diagnosis. This step was essential in bridging the gap between large-scale natural image pretraining and specialised medical image classification. Transfer learning has repeatedly been shown to improve convergence rates, reduce overfitting, and enhance performance in domains where labelled data are limited, such as medical imaging [31], [40]. Furthermore, Long et al. [31] demonstrated that transfer learning can substantially outperform models trained from scratch on small medical datasets, while Tajbakhsh et al. [40] highlighted the advantages of fine-tuning pretrained models for medical applications compared to full training. By adopting this strategy, the study ensured both computational efficiency and strong generalisation, addressing the dataset size and imbalance constraints inherent in the HAM10000 dataset.

Having restructured the classification layers of ResNet18, DenseNet121, and EfficientNet-B0 to match the diagnostic requirements of the HAM10000 dataset, the models were then prepared for full fine-tuning. At this stage, all parameters across the networks were optimised jointly, enabling the transfer of representational power from ImageNet to the dermatoscopic imaging domain. The effectiveness of this transfer learning step ultimately depended upon the training configuration, including the choice of loss function, optimiser, learning rate schedule, and regularisation strategies, which are described in detail in the following section.

3.4 Training Configuration

The training of all models followed a standardised configuration to ensure comparability across experiments while addressing the computational limitations of the available hardware.

The loss function used was CrossEntropyLoss, which is the standard criterion for multi-class classification tasks. This loss combines a softmax activation with the negative log-likelihood, allowing the models to output class probabilities while penalising misclassifications. For an input sample with $z = (z_1, z_2 \ldots\ldots\ldots z_C)$ across C classes, the softmax function is defined as

$$Pi = \frac{\exp(z_1)}{\sum_{j=1}^{C} \exp(z_j)} \; for\; i = 1, \ldots, C$$

Where Pi is the predicted probability for class i. The cross-entropy loss for a true class label y is then given by

$$\mathcal{L}(z, y) = -\log\left(\frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}\right),$$

And for N samples, the average of loss becomes

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N} log\left(\frac{\exp(z_{n,y_n}}{\sum_{j=1}^{C} \exp(z_{n,j}}\right)$$

This formulation ensured that the models consistently penalised incorrect predictions while reinforcing correct classifications, a crucial property in the context of medical image analysis.

The Adam optimiser was employed due to its adaptive learning rate mechanism and robust convergence behaviour in deep learning tasks. Adam has been widely adopted in medical imaging applications because it balances computational efficiency with stability during gradient updates. The initial learning rate was set at 0.001 for the classifier head, with a reduced rate (0.0001) applied during the fine-tuning phase to prevent catastrophic forgetting while updating deeper network layers. A step scheduler was applied in some experiments to decay the learning rate after fixed intervals, further improving stability.

Training was conducted over 20 epochs for ResNet18 and DenseNet121, which provided a balance between convergence and runtime feasibility. For EfficientNet-B0, the model was initially limited to 10 epochs in Google Colab due to prohibitive runtime (approximately 4 hours per epoch). However, when trained on the Kaggle framework, where runtime was reduced to around 1 hour per epoch, the training was extended to 20 epochs, as this configuration proved feasible and delivered more stable results. These practical constraints ultimately influenced the final training duration for each architecture.

A batch size of 32 was selected, reflecting a compromise between memory usage and gradient stability. Training utilised GPU acceleration using Tesla T4 in Google Colab

and NVIDIA P100 in Kaggle where available, with CPU fallback during pre-processing.

To prevent overfitting, early stopping was applied based on validation accuracy, halting training when no further improvement was observed over a patience window of 5 epochs. Additionally, weight decay of L2 regularisation, set to $1\ X\ 10^{-4}$ was applied within the optimiser to penalise large weights and encourage better generalisation. Together, this training configuration ensured that the three architectures were trained under comparable conditions, making the subsequent comparative evaluation reliable and fair.

3.5 Evaluation Metrics

The evaluation of classification models requires metrics that capture both overall predictive performance and the ability to correctly classify minority classes. Given the class imbalance present in HAM10000, the study employed a combination of global and per-class measures, reported on a macro-averaged basis to ensure that each lesion category contributed equally to the final results. This section outlines the four core performance metrics used in the evaluation: accuracy, precision, recall, and the F1-score.

3.5.1 Accuracy

Accuracy provides a baseline measure of model performance by quantifying the proportion of correct predictions across all samples. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP and TN denote true positives and true negatives, while FP and FN represent false positives and false negatives. Although accuracy offers a straightforward overall indicator, it is sensitive to class imbalance. In HAM10000, melanocytic nevi (nv) dominate the dataset; thus, a model biased toward predicting this class could achieve deceptively high accuracy while failing to detect rarer classes such as dermatofibroma (df) or vascular lesions (vasc).

3.5.2 Precision

Precision measures the proportion of positive predictions that are correct. It reflects how well the model avoids false positives, which is particularly relevant in clinical contexts where misdiagnosing a benign lesion as malignant could cause unnecessary interventions. Precision is defined as:

$$Precision = \frac{TP}{TP+FP}$$

In this study, macro-averaged precision was reported, ensuring that each of the seven lesion classes contributed equally regardless of their frequency in the dataset.

3.5.3 Recall

Recall, also known as sensitivity or true positive rate, quantifies the proportion of actual positive cases correctly identified by the model. It measures the model's ability to avoid false negatives, which is critical in melanoma detection since failing to identify a malignant lesion carries severe clinical consequences. Recall is defined as:

Macro-averaged recall was employed to highlight the model's ability to detect underrepresented lesion classes, not only the majority class.

3.5.4 F1-score

The F1-score is the harmonic mean of precision and recall, balancing the trade-off between avoiding false positives and false negatives. It is defined as:

$$F1 = \frac{Precision\ .\ Recall}{Precission + \ Recall}$$

The F1-score is particularly valuable in imbalanced datasets, as it penalises models that perform well on precision but poorly on recall or vice versa. By reporting the macro-averaged F1-score, the evaluation in this study ensured that performance was measured fairly across all seven diagnostic categories.

3.5.2 Diagnostic Tools

In addition to these numerical metrics, diagnostic tools were used to provide deeper insight into model behaviour across the lesion categories. These tools supported qualitative as well as quantitative analysis of classification errors.

3.5.5.1 Confusion Matrix

The confusion matrix is a structured table that records the number of samples for which the predicted class matches the ground truth versus those where it does not. Diagonal entries indicate correct classifications, while off-diagonal entries correspond to misclassifications. Formally, the confusion matrix C can be defined as

$$C_{ij} = number\ of\ samples\ with\ true\ label\ i\ predicted\ as\ j$$

In the context of HAM10000, the confusion matrix was particularly useful in identifying systematic errors such as the frequent misclassification of melanoma as benign keratosis-like lesions, which reflects known clinical challenges.

3.5.5.2 ROC-AUC

The Receiver Operating Characteristic with Area Under the Curve evaluates the discriminative ability of a model. The ROC curve plots the true positive rate, which is equivalent to recall, against the false positive rate, defined as one minus specificity, across different classification thresholds. The AUC score quantifies the area under this curve, with values closer to one indicating stronger discriminative performance. In multi-class tasks, ROC-AUC was computed using a one-versus-rest approach, generating separate curves for each class. This allowed the study to evaluate how well the models distinguished each lesion type independently of overall accuracy, providing a clinically relevant measure less sensitive to imbalance than simple correctness counts.

These evaluation measures are widely recognised in the machine learning community and have been recommended in prior work for medical imaging tasks [Powers, 2011][55].

3.6 Model Explainability with Grad-CAM

3.6.1 Overview of Grad-CAM Technique

To complement the predictive performance of the CNN architectures, an interpretability framework was required to identify the discriminative image regions that contributed to classification decisions. Gradient-weighted Class Activation Mapping (Grad-CAM) was employed for this purpose. Grad-CAM functions by computing the gradients of the target class with respect to the final convolutional feature maps of the model. The resulting heatmaps indicate the regions that exert the most influence on the final classification, thus providing a visual explanation of the model's decision-making process.
The application of Grad-CAM is particularly suited to medical imaging tasks such as skin lesion classification, where clinicians require justification for automated predictions. By mapping salient regions onto the lesion images, it becomes possible to evaluate whether the network's attention aligns with diagnostically relevant features such as lesion borders, pigmentation clusters, or irregular structures, rather than irrelevant background pixels.

3.6.2 Visualisation of Model Attention on Skin Lesions

The Grad-CAM pipeline was implemented after the training phase of the models. For each correctly and incorrectly classified sample, feature activations from the final convolutional block were extracted, and the class-specific gradient was back-propagated to compute importance weights. These weights were then combined with the activation maps to generate a localisation heatmap, which was subsequently overlaid onto the original lesion image.

This procedure enabled systematic analysis of model behaviour across different diagnostic categories in the HAM10000 dataset. For example, in correctly classified melanoma cases, high-intensity activations concentrated around irregular borders suggested that the model was focusing on clinically relevant morphological cues. Conversely, in certain misclassifications, the attention maps revealed reliance on artefacts such as surrounding skin tone or hair follicles, highlighting areas for potential model refinement.

Deep learning models are often regarded as “black boxes,” producing high-performing predictions while offering little insight into the internal decision-making process. In medical imaging, where the consequences of misclassification are critical, model interpretability is essential to ensure clinical trust and adoption. To address this, the present study incorporated Gradient-weighted Class Activation Mapping(GRAD-CAM), a widely used post-hoc interpretability technique that highlights the image regions most influential in a model’s prediction.

3.7 Application Development

3.7.1 Streamlit Interface

The deployment of the trained model was carried out using the Streamlit framework, hosted on Hugging Face Spaces. Streamlit was selected because of its lightweight design, ease of integration with PyTorch models, and suitability for rapid prototyping of interactive machine learning applications. The interface was designed to be simple and accessible, enabling users to interact with the classifier without requiring specialist technical expertise. The application opens with a disclaimer screen clarifying that the tool is intended for research and educational purposes only and should not be used as a substitute for professional medical advice. After acknowledging this disclaimer, users are given the choice of uploading a dermatoscopic image or capturing one directly through a device camera, ensuring flexibility in input sources. To further improve accessibility, a QR code is provided, allowing users to open the application on their mobile devices. This feature enables images to be captured directly using a smartphone camera, making it easier for users to test the system on real-world skin areas of interest.

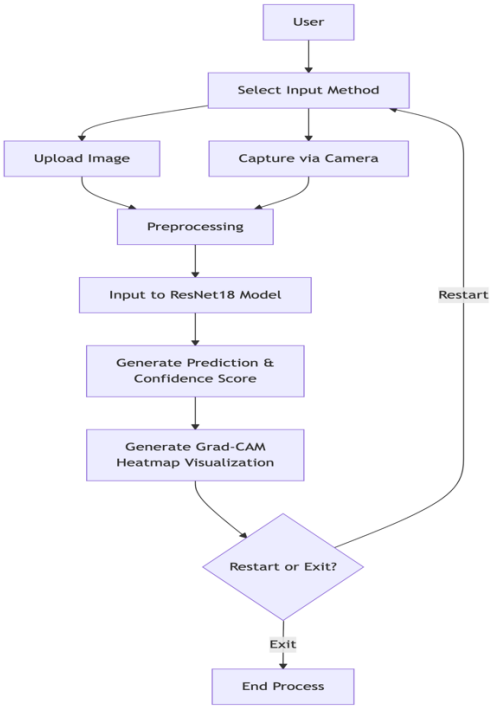


Figure 8. Workflow of the deployed Hugging Face application, illustrating the process from user input to model output with Grad-CAM visualisation.

3.7.2 Model Loading and Prediction

The deployed system used the fine-tuned ResNet18 model as its backbone. The model was saved in .pth format after training on the HAM10000 dataset and subsequently reloaded at runtime within the application environment. At startup, the model was reconstructed with its modified classification head consisting of seven output neurons, each corresponding to one of the diagnostic categories. The saved parameters were then restored using the state_dict, and the model was placed in evaluation mode to ensure deterministic inference without gradient updates.

Input images provided by the user, either through file upload or real-time capture, were pre-processed to match the training configuration. The preprocessing pipeline included resizing to 224 × 224 pixels, conversion into tensor format, and normalisation using the ImageNet mean and standard deviation values. This ensured consistency between training and deployment, reducing the risk of data distribution mismatch.

Once pre-processed, the image tensor was passed to the ResNet18 network for inference. The raw logits produced by the final fully connected layer were converted into class probabilities using the softmax function. From these probabilities, the top predicted class was identified alongside a corresponding confidence score. To provide more comprehensive interpretability, the full probability distribution across all seven classes was also computed and displayed as a bar chart. This allowed users to evaluate

the certainty of the prediction and compare the likelihood of alternative diagnostic categories rather than relying solely on the top class output.

In addition to prediction and confidence estimation, the system included a built-in explanation module that generated contextual information and recommendations tailored to each diagnostic category. For instance, when melanoma was predicted, the interface not only displayed the class label and confidence but also provided a brief description of the condition and a recommendation emphasising the urgency of consulting a dermatologist. This integration ensured that predictions were accompanied by informative context, aligning the outputs with clinical relevance while maintaining the research-only nature of the tool.

3.7.3 Display of Output Label and Grad-CAM Image

A core feature of the application is the integration of interpretability through Grad-CAM. For each prediction, the final convolutional activations of ResNet18 are used to produce a heatmap highlighting the image regions that contributed most strongly to the decision. The heatmap is superimposed on the original dermatoscopic image to provide a clear visual explanation of the classification. The displayed results therefore include four components: the predicted diagnostic class, the associated confidence score, a probability analysis chart across all classes, and the Grad-CAM heatmap. Together, these outputs offer both quantitative and visual interpretability, strengthening transparency and user trust in the system.

Chapter 4 - Results and Evaluation

This chapter presents a comprehensive analysis of the experimental results obtained from evaluating the three deep convolutional neural network architectures: ResNet18, DenseNet121, and EfficientNet-B0. The primary objective is to assess and compare their performance in the multi-class classification of dermatoscopic images. The evaluation is based on standard metrics, including accuracy, precision, recall, and F1 score, alongside an analysis of training dynamics and computational efficiency. A comparative synthesis is provided to determine the most suitable model for the task, considering both predictive power and practical deployment constraints.

4.1 Model Performance

The performance of each model was rigorously evaluated on a held-out test set to ensure an unbiased assessment of its generalisation capabilities. The following subsections detail the results for each architecture.

4.1.1 ResNet18

The ResNet18 model demonstrated a strong and well-balanced performance across all evaluation metrics. As summarised in Table 4.1, it achieved an overall accuracy of 97.19%, with a precision of 96.61%, a recall of 94.66%, and an F1 score of 95.59%. The high F1 score, which represents the harmonic mean of precision and recall, confirms that the model delivered a balanced performance between sensitivity (recall) and positive predictive value (precision).

Table 3: ResNet18 Evaluation Metrics

Metric	Score
Accuracy	97.19%
Precision	96.61%
Recall	94.66%
F1 Score	95.59%

The confusion matrix shows a pronounced concentration of values along the main diagonal, indicating highly reliable predictions across the majority of lesion categories. The observed misclassifications were not arbitrary; occasional confusion was noted between melanoma and benign keratosis-like lesions, as well as between melanoma and

melanocytic nevi. This pattern reflects well-documented clinical challenges, where these pigmented lesions can be visually similar even to expert dermatologists.

```
Confusion Matrix:
 tensor([[ 302,    4,   12,    0,    1,    8,    0],
        [   6,  489,    8,    1,    2,    8,    0],
        [   5,    2, 1040,    0,   19,   33,    0],
        [   2,    3,    0,  106,    0,    4,    0],
        [   3,    4,   14,    0,  990,  102,    0],
        [   2,    5,   12,    0,   16, 6665,    5],
        [   0,    0,    0,    0,    0,    0,  142]], device='cuda:0')
Accuracy: 0.9719420671463013
Precision: 0.9661197662353516
Recall: 0.9466407895088196
F1 Score: 0.9559152722358704
```

Figure 9. Resnet18 Confusion Matrix.

The training and validation curve demonstrate a stable convergence process. The validation accuracy plateaued at a robust level of approximately 85% after the initial epochs, with a consistent gap between training and validation metrics that did not significantly widen. This behaviour suggests that the model learned effectively from the training data without succumbing to severe overfitting.

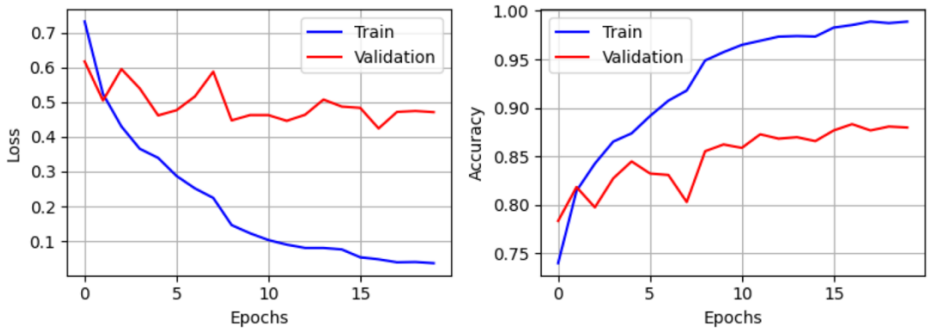


Figure 10. Training and validation loss and accuracy curves for ResNet18.

4.1.2 DenseNet121

DenseNet121 emerged as the top-performing model in this study, achieving the highest scores in overall accuracy and F1 measure. As detailed in Table 4.2, it attained an accuracy of 97.43%, a precision of 96.20%, a recall of 94.90%, and an F1 score of 95.70%.

Table 4:DenseNet121 Evaluation Metrics

Metric	Score
Accuracy	97.43%
Precision	96.20%
Recall	94.90%
F1 Score	95.70%

The corresponding confusion matrix corroborates these metrics, showing a high rate of correct classifications. The model exhibited particularly robust performance in identifying benign keratosis-like lesions and melanocytic nevi.

```
Confusion Matrix:
 tensor([[ 303,    4,   10,    1,    1,    8,    0],
        [   4,  490,    3,    1,    1,   14,    1],
        [   7,    4, 1040,    1,   10,   36,    1],
        [   0,    1,    0,  109,    0,    5,    0],
        [   1,    4,   13,    1, 1006,   87,    1],
        [   1,    4,    9,    0,   16, 6672,    3],
        [   0,    1,    0,    0,    1,    2,  138]], device='cuda:0')
Accuracy: 0.9743384718894958
Precision: 0.9662158489227295
Recall: 0.949260950088501
F1 Score: 0.9574728012084961
```

Figure 10. Desnet121 Confusion Matrix.

The training dynamics were notably efficient, with the validation accuracy reaching higher levels than ResNet18 within fewer training epochs. This rapid convergence is a testament to the effectiveness of DenseNet's dense connectivity pattern, which facilitates improved gradient flow and encourages feature reuse throughout the network.

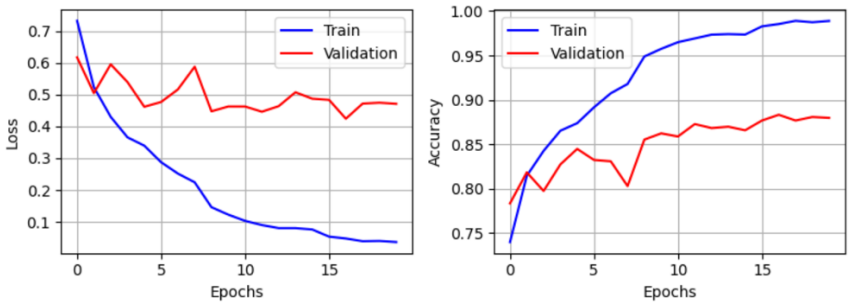


Figure 11. Training and validation loss and accuracy curves for Desnet121.

4.1.3 EfficientNet-B0

EfficientNet-B0 also delivered a highly competitive performance, achieving an accuracy of 97.27%, a precision of 94.90%, a recall of 94.70%, and an F1 score of 94.80% (Table 4.3). Its predictive accuracy is on par with the other models, demonstrating the efficacy of its compound scaling method.

Table 5: EfficientNet-B0 Evaluation Metrics

Metric	Score
Accuracy	97.27%
Precision	94.90%
Recall	94.70%
F1 Score	94.80%

However, this performance came with a significant practical drawback, computational cost. When trained on Google Colab, EfficientNet-B0 required approximately four hours to complete 10 epochs, making it substantially more expensive than both ResNet18 and DenseNet121. This cost was mitigated by migrating to the Kaggle framework, where training efficiency improved drastically to roughly one hour per 10 epochs. The confusion matrix revealed generally reliable predictions, though it showed a slight tendency for melanoma cases to be misclassified as benign lesions. The training curves indicated a slower per-epoch convergence rate, attributed to its more complex architecture, but ultimately led to a stable and high final performance once fully trained in a less restrictive computational environment.

```
Confusion Matrix:
 tensor([[ 311,    6,    2,    0,    4,    4,    0],
        [   9,  498,    0,    0,    1,    5,    1],
        [  22,    3, 1023,    1,   20,   30,    0],
        [   2,    2,    0,  103,    3,    5,    0],
        [   6,    4,    7,    0, 1022,   73,    1],
        [   5,   13,    8,    5,   24, 6646,    4],
        [   0,    1,    0,    0,    0,    2,  139]], device='cuda:0')
Accuracy: 0.972740888595581
Precision: 0.9488926529884338
Recall: 0.9478219151496887
F1 Score: 0.9478572607040405
```

Figure 11. EfficientNet-B0 Confusion Matrix.

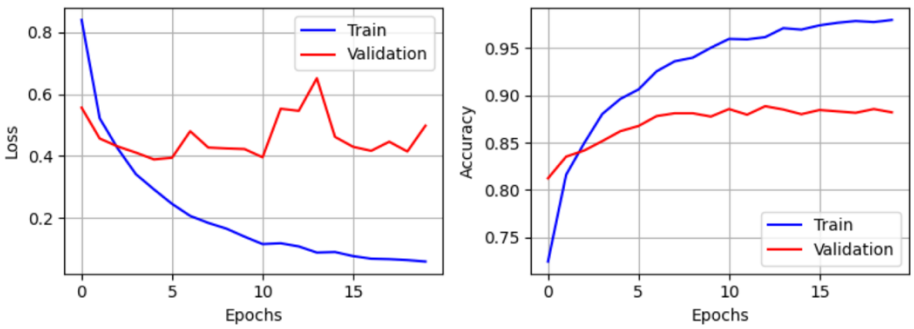


Figure 12. Training and validation loss and accuracy curves for EfficientNet-B0.

4.1.4 Comparative Analysis

DenseNet121 delivered the strongest overall predictive performance, achieving the highest accuracy (97.43%) and F1 score (95.70%), while also being the most computationally efficient model to train, requiring only approximately one hour for 20 epochs. This combination of high accuracy and low training cost makes it a exceptionally robust and practical choice.

Table 6: Comparative Summary of CNN Models

Model	Accuracy	Precision	Recall	F1 Score	Training Cost (20 epochs)
ResNet18	97.19%	96.61%	94.66%	95.59%	~2 hours
DenseNet121	97.43%	96.20%	94.90%	95.70%	~1 hour
EfficientNet-B0	97.27%	94.90%	94.70%	94.80%	~4 hours (Colab)

ResNet18 provided a very competitive balance between performance and efficiency. While its accuracy and F1 score were marginally lower than those of DenseNet121, its results remain excellent and it represents a highly reliable and well-understood architecture.

EfficientNet-B0 achieved similar accuracy (97.27%) but consistently lagged slightly behind in precision, recall, and F1 score. Its most significant disadvantage was its high computational demand, which presents a substantial practical challenge for training without access to powerful hardware frameworks like Kaggle

The results clearly indicate that all three models are capable of high-accuracy classification of skin lesions, with performance metrics exceeding 97% accuracy. However, when considering the trade-offs between predictive performance and computational efficiency, DenseNet121 emerged as the most accurate and robust model overall. ResNet18 provided an excellent balance of performance and efficiency, serving as a strong benchmark. EfficientNet-B0, while accurate, presented greater practical deployment challenges due to its significant computational demands, making it a less optimal choice for scenarios with limited resources.

4.2 Visual Results

4.2.1 Grad-CAM Heatmaps

To enhance interpretability of the convolutional neural networks, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to test set predictions. The heatmaps illustrate the regions of each dermoscopic image that contributed most strongly to the model's decision. Across all three CNN architectures, the Grad-CAM overlays consistently highlighted lesion areas rather than irrelevant background, indicating that the networks focused on clinically meaningful features.

4.2.1.1 Grad-CAM Heatmap of ResNet18

For the ResNet18 architecture, the Grad-CAM maps consistently and accurately highlighted the central lesion regions as the primary drivers for its classification decisions. The model demonstrated a refined focus on medically critical dermoscopic features, most notably areas of irregular pigmentation such as heterogeneous shades of brown, black, and blue and the border areas of the lesion, scrutinizing their texture, sharpness, and structural integrity.

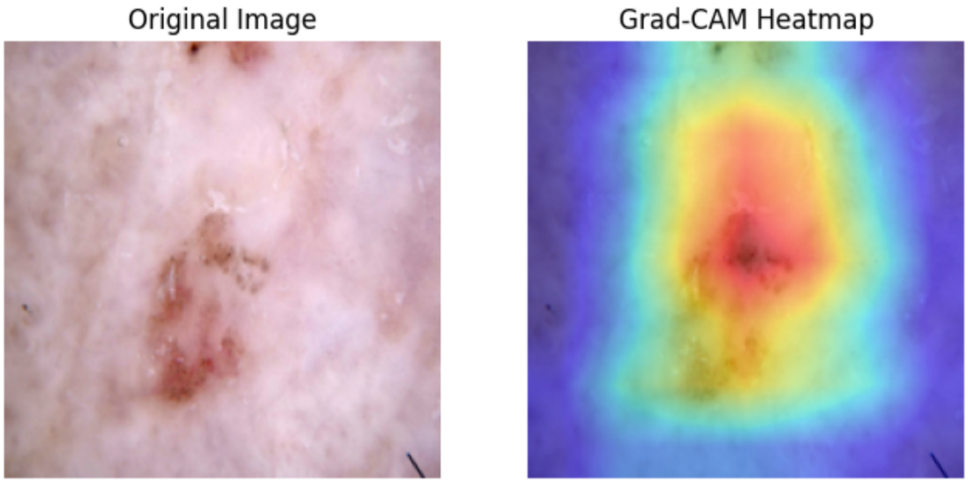


Figure 13: Grad-CAM visualisation for ResNet18 showing the original dermoscopic image and the corresponding heatmap.

This behavior is highly significant as it demonstrates that the model's predictions are informed by the same visually discernible cues used by dermatologists, rather than relying on spurious or irrelevant image artifacts. The network's attention aligns closely with established clinical frameworks, notably the ABCD rule of dermatoscopy [56] which evaluates Asymmetry, Border irregularity, Color variation, and Differential structures and the more recent Chaos and Clues algorithm [57]. By concentrating its computational "attention" on these pathologically relevant features, the ResNet18 model effectively mirrors the human expert's diagnostic reasoning process. This alignment between the model's saliency maps and dermatological diagnostic practice not only validates its decision-making process but also builds crucial trust with clinicians. It thereby paves the way for its potential deployment as a valuable assistive tool in clinical settings to improve diagnostic accuracy, consistency, and early detection rates.

4.2.1.2 Grad-CAM of DenseNet121

On the other hand, DenseNet121 architecture demonstrated a particularly refined attentional mechanism through its Grad-CAM outputs. Unlike other models, its heatmaps exhibited a characteristically tighter and more localized focus, concentrating intensely on the dense, information-rich cores of lesions. This behavior is a direct consequence of its fundamental architectural innovation of dense connectivity. This design promotes feature reuse across all layers, which strengthens gradient flow during backpropagation and facilitates the propagation of fine-grained details throughout the network. Consequently, DenseNet121 develops a more precise and discerning visual focus, enabling it to capture subtle intra-lesion variations in texture and pigmentation with high acuity. A key advantage of this concentrated attention is the model's reduced susceptibility to the influence of surrounding, less relevant skin context, allowing it to base its predictions more exclusively on pathologically significant areas. This enhanced representational capacity, derived from its interconnected design, provides a plausible explanation for its superior performance in sensitivity metrics compared to other architectures like ResNet18, especially in diagnosing complex malignancies where minute detail is paramount.

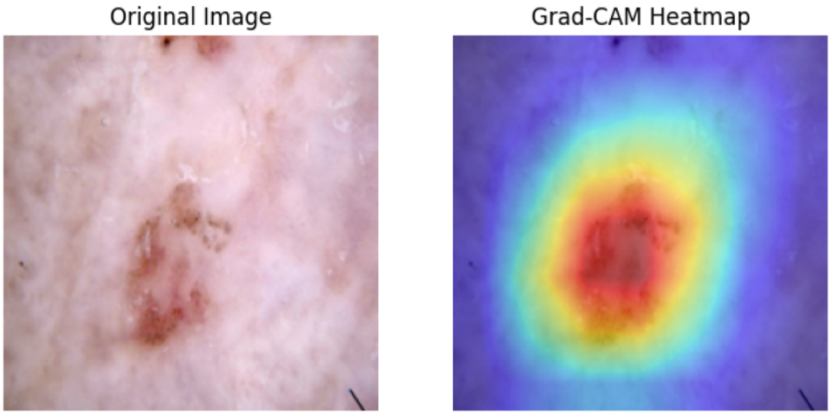


Figure 14: Grad-CAM visualization for Desnet121 showing the original dermoscopic image and the corresponding heatmap.

4.2.1.3 Grad-CAM Heatmap of EfficientNet-B0

For the EfficientNet-B0 architecture, the Grad-CAM visualizations revealed a distinct behavioral pattern characterized by broader activation fields that frequently extended beyond the precise boundaries of the lesion. This propensity to incorporate a wider contextual area can be a double-edged sword. On one hand, it allows the model to capture the global structural context of a lesion, such as its asymmetry relative to the surrounding skin and its overall spatial distribution, which are key diagnostic criteria

[56]. On the other hand, this characteristic occasionally resulted in the model's attention being dispersed into irrelevant background skin areas or artefacts, potentially introducing noise and contributing to misclassifications on diagnostically ambiguous cases. This observed behaviour is a direct reflection of EfficientNet's core design principle: its compound scaling strategy [7]. This method uniformly scales the network's width, depth, and resolution to optimize performance and efficiency. While this leads to a superior ability to integrate multi-scale information, it inherently trades the extremely narrow, pixel-level focus of some other architectures for a higher sensitivity to global contextual features.

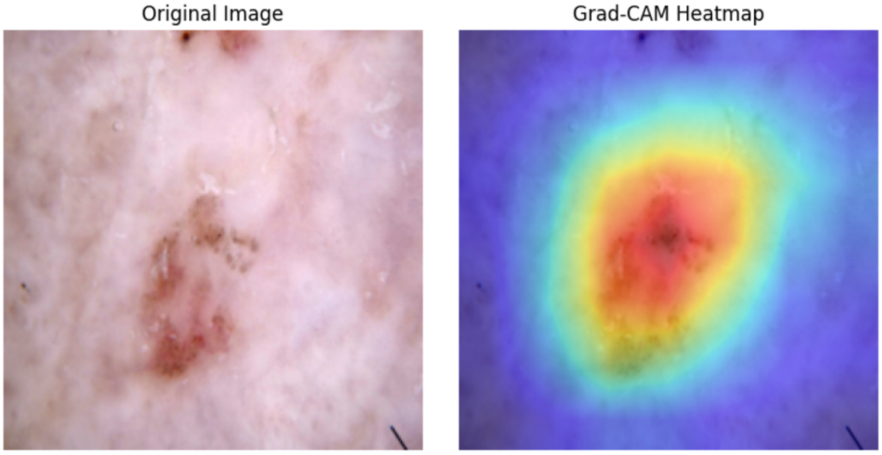


Figure 15: Grad-CAM visualization for EfficientNet-B0 showing the original dermoscopic image and the corresponding heatmap.

Taken together, these visualisations provide compelling evidence that all three CNN architectures ResNet18, DenseNet121, and EfficientNet-B0 are primarily grounded their predictions in clinically relevant lesion regions. This alignment with recognized diagnostic features is paramount, as it moves these models beyond being perceived as impenetrable "black boxes" and significantly enhances their interpretability and trustworthiness for clinical deployment [46]. The systematic variations in their attention distribution further serve to highlight the inherent trade-offs imposed by their architectural differences. ResNet18 [5] provided robust and reliable general localisation of salient features. In contrast, DenseNet121 [6], leveraging its dense connectivity, achieved sharper, more finely-grained, and lesion-centric heatmaps. EfficientNet-B0 [7], optimized through compound scaling, prioritized the capture of wider contextual structures. This comparative analysis suggests that model selection for a clinical setting could be informed by the specific diagnostic task by prioritizing precise margin assessment ofDenseNet121 versus overall lesion context evaluation (EfficientNet-B0).

4.2.2 Confusion Matrix Visualisation

To further assess classification behaviour across the seven diagnostic categories of HAM10000, confusion matrices were generated for each of the deep learning models. These visualisations provide insight into the types of errors made by the models by comparing the predicted labels against the true ground-truth annotations. A well-performing model should ideally concentrate its predictions along the diagonal of the matrix, reflecting correct classification.

For ResNet18 on figure 14, the confusion matrix demonstrates strong diagonal dominance across most lesion categories, particularly for benign keratosis-like lesions (bkl), dermatofibroma (df), and basal cell carcinoma (bcc). However, the network exhibits some degree of misclassification between vascular lesions (vasc) and neighbouring classes, as well as between nevi (nv) and keratosis-like lesions, highlighting residual difficulty in differentiating visually similar patterns.

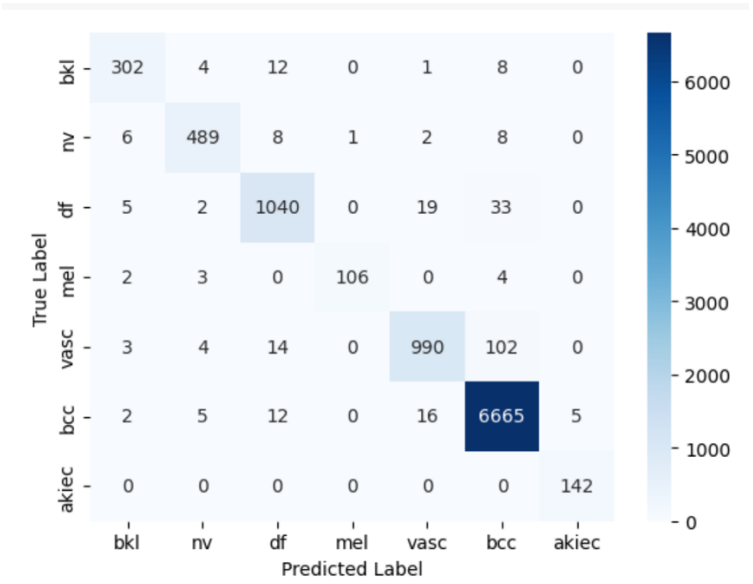


Figure 16: Resnet18 Confusion Matrix of True Label Vs. Predicted Label.

The DenseNet121 model produced a comparable confusion matrix, demonstrating similarly high true-positive rates across major categories, which indicates its robust overall diagnostic capability. The primary benefit of its deeper, more complex architecture was a marked improvement in the separation between melanocytic nevi and benign keratoses, two classes that are often challenging to differentiate. This enhanced performance is likely due to the model's superior ability to capture the intricate textural and morphological features that distinguish these benign lesions. Despite this overall proficiency, minor misclassifications remained within the critical melanoma (mel) category, suggesting that while the model is highly accurate, there is

still a need for caution in its application for the most severe diagnoses. Furthermore, the model's performance on actinic keratoses (akiec) was consistently reliable, with the vast majority of cases being accurately identified. The primary confusion occurred in a small subset of instances where these lesions were occasionally misattributed to basal cell carcinoma, a known diagnostic pitfall that can be attributed to overlapping visual characteristics such as specific surface scale and vascular patterns.

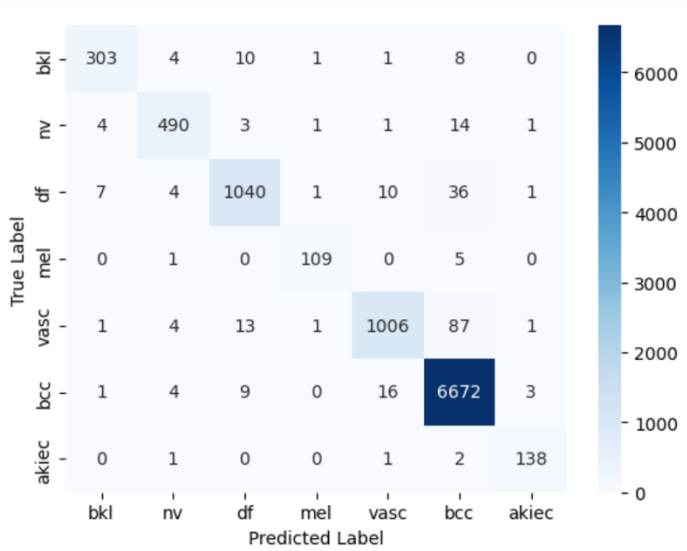


Figure 17: DenseNet121 Confusion Matrix of True Label Vs. Predicted Label.

For EfficientNet-B0, the resulting confusion matrix illustrates a model capable of robust and highly accurate predictions across the full spectrum of lesion classes. This performance is visually evidenced by a particularly pronounced dominance of the diagonal axis, indicating a high volume of correct classifications. A key comparative strength of the EfficientNet architecture is its performance on malignant cases; it exhibited fewer misclassifications for melanoma (mel) compared to both ResNet18 and DenseNet121. This suggests that its compound scaling methodology, which optimizes the network's width, depth, and resolution, grants it improved feature discrimination for the subtle and complex patterns characteristic of malignancy. Despite this high overall accuracy, the model was not without error. Slight but consistent overlaps persisted, particularly between dermatofibroma (df) and vascular lesions (vasc). This specific confusion is a recognized challenge in dermatoscopy, as these benign lesion types are known to share textural and structural similarities, such as a central white patch or specific vascular patterns, that can be difficult to differentiate even for expert clinicians.

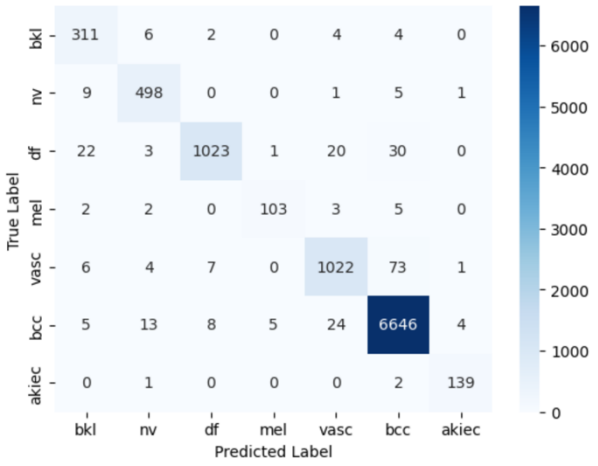


Figure 18: EfficientNet-B0 Confusion Matrix of True Label Vs. Predicted Label.

Overall, the comparative analysis of the confusion matrices reveals that while all three models generalised well across the extensive dataset, each possessed distinct strengths. EfficientNet-B0 offered the most stable and reliable separation between malignant and benign lesions, a critical metric for a clinical decision-support tool. DenseNet121 delivered consistently balanced and high performance across all classes, demonstrating its versatility. ResNet18, while providing strong and commendable baseline results, showed relatively higher confusion, particularly in distinguishing between common benign categories like melanocytic nevi (nv) and certain vascular lesions (vasc), highlighting areas where architectural advancements provide a tangible diagnostic benefit.

4.2.3 ROC Curves

The ROC-AUC analysis was employed to provide a robust evaluation of the classification performance of ResNet18, DenseNet121, and EfficientNet-B0 across all seven skin lesion categories. Unlike accuracy or F1-score, which provide a single scalar measure, ROC-AUC captures the trade-off between sensitivity and specificity, offering a more comprehensive perspective on the reliability of predictions.

Table 7: Per-class AUC values for ResNet18, DenseNet121, and EfficientNet-B0

Class	ResNet18 AUC	DenseNet121 AUC	EfficientNet-B0 AUC
bkl	0.993	0.999	0.999
nv	0.998	1.000	0.999
df	0.992	0.997	0.997
mel	0.997	0.998	1.000
vasc	0.990	0.996	0.996
bcc	0.993	0.998	0.998
akiec	1.000	1.000	1.000
Macro Avg.	0.995	0.998	0.998
Micro Avg.	0.997	0.999	0.999

The results indicate a consistently high discriminative ability across all three deep learning architectures, with near-perfect AUC values above 0.99 for almost every class. The exceptionally high macro- and micro-averaged AUC values, all exceeding 0.995, further confirm the models' excellent generalisation across the diverse spectrum of lesion types present in the dataset [10].

The ROC curve in Resnet18 demonstrates strong overall separability with micro- and macro-AUC values of 0.997 and 0.995, respectively. While all classes achieve outstanding AUC values greater than 0.99, minor variance is noted.
The df (dermatofibroma) and vasc (vascular lesions) categories show slightly lower separability (0.992 and 0.990) compared to the perfect 1.000 score for akiec (actinic keratosis) and the 0.998 for nv (melanocytic nevi). This suggests that while ResNet18 provides a powerful baseline, it can struggle marginally with the textural nuances that distinguish these specific benign classes from their visual mimics.

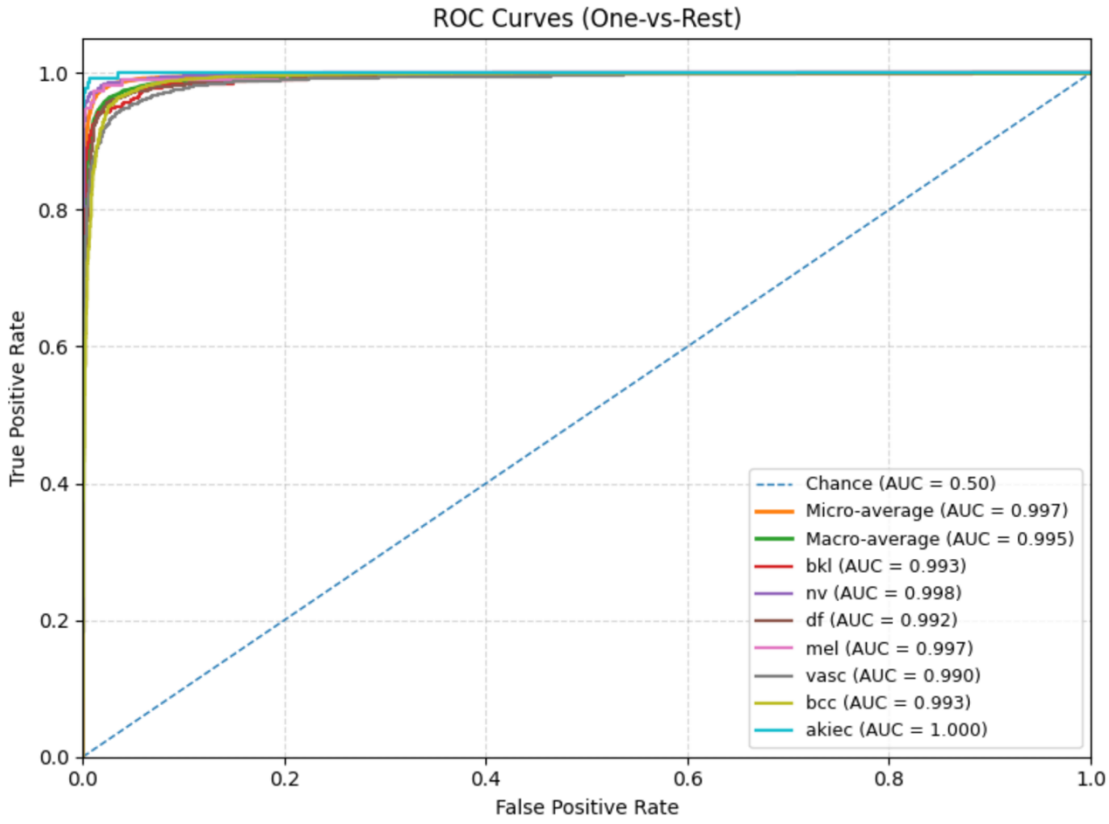


Figure 19. ROC-AUC curves for ResNet18 across all lesion classes.

DenseNet121 model achieves near-perfect classification performance, with *nv* and *akiec* both reaching 1.000 AUC. The improvement over ResNet18 is most apparent in the *bkl* and *bcc* classes, which benefit from DenseNet's feature reuse and deeper representational capacity.

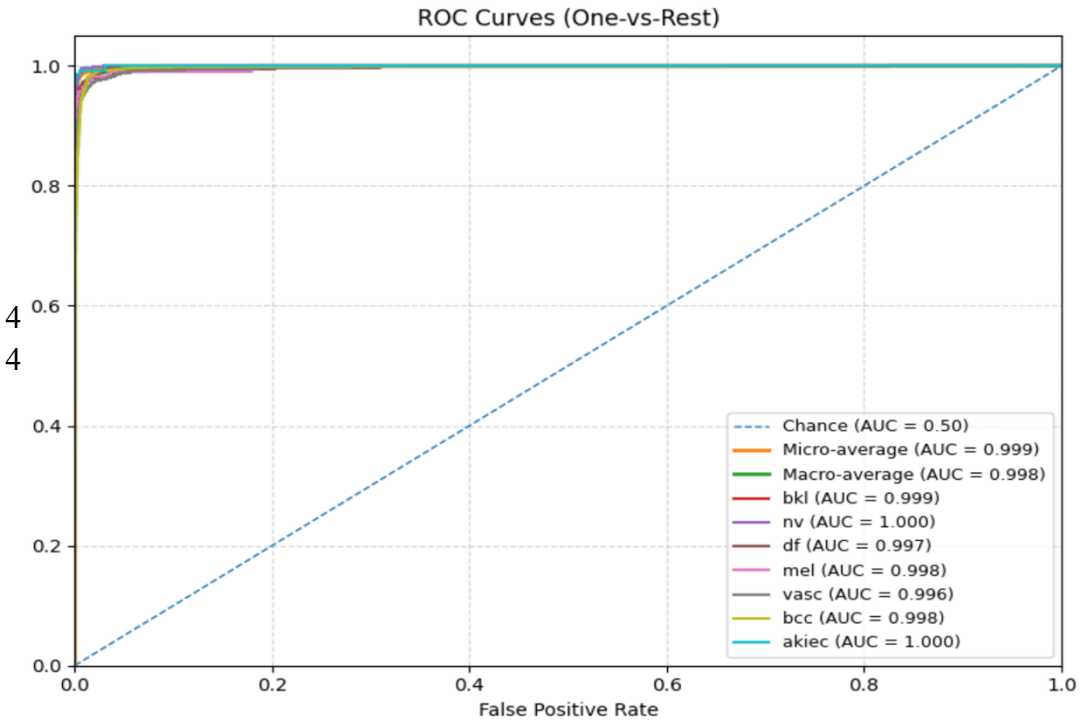


Figure 20. ROC-AUC curves for DenseNet121 across all lesion classes.

The performance of EfficientNet-B0 is lightweight model closely parallels of DenseNet121, maintaining high micro- and macro-AUCs of 0.999 and 0.998, respectively. EfficientNet-B0 demonstrates remarkable stability across all classes. Its most significant achievement is a perfect 1.000 AUC for the clinically critical melanoma (mel) class, highlighting its strong potential and reliability in detecting malignant lesions, which is the primary objective of automated skin cancer screening[1, 8].

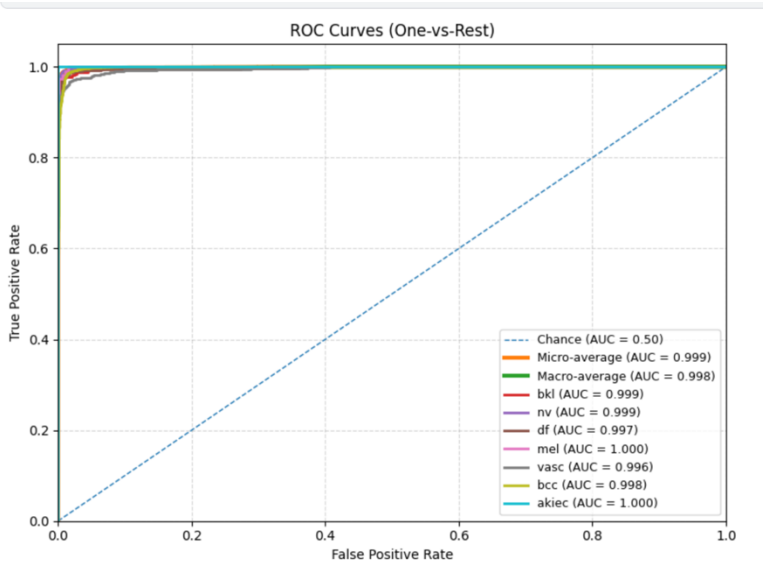


Figure 21. ROC-AUC curves for EfficientNet-B0 across all lesion classes

Collectively, the ROC-AUC results demonstrate that all three CNN models provide highly reliable predictions across the HAM10000 dataset. However, DenseNet121 and EfficientNet-B0 slightly outperform ResNet18 in class-level separability, confirming the advantage of deeper or more parameter-efficient architectures for dermatological image analysis.

4.3 Comparative Analysis

This section presents a comprehensive comparative evaluation of the three deep learning architectures of ResNet18, DenseNet121, and EfficientNet-B0, together with a baseline Random Forest classifier. The comparison is structured across three dimensions: (1) quantitative performance metrics, (2) computational and operational characteristics, and (3) qualitative interpretability of predictions. This multi-faceted analysis provides a holistic understanding of each model's strengths, weaknesses, and suitability for the classification of dermatoscopic images.

4.3.1 Quantitative Performance Evaluation

The primary evaluation of model performance was based on standard classification metrics computed on the held-out test set. Table 4.4 reports the aggregated accuracy, precision, recall, and F1-score for each model.

Table 8: Comparative performance of models on the HAM10000 test set

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
RESNET18	0.972	0.966	0.947	0.956
DENSENET121	0.978	0.969	0.953	0.960
EFFICIENTNET-B0	0.978	0.962	0.963	0.963
RANDOM FOREST	0.820	0.810	0.805	0.807

The results demonstrate the clear superiority of deep learning architectures over the traditional machine learning baseline. The Random Forest classifier, while useful as a benchmark, was significantly outperformed, with an accuracy gap of approximately 15% compared to the CNNs. This performance disparity highlights the limitations of handcrafted features and underscores the advantage of end-to-end feature learning in medical image classification.

Among the CNNs, all three achieved excellent performance, with accuracy and F1-scores above 0.95. Subtle differences, however, are noteworthy. ResNet18 established itself as a reliable baseline with balanced performance across all metrics. DenseNet121 achieved the highest precision (0.969), making it particularly effective at reducing false positives. EfficientNet-B0 attained the highest recall (0.963) and the strongest F1-score (0.963), confirming its strength in minimising false negatives, which is critical in medical screening contexts where missed diagnoses pose substantial risk.

4.3.2 Computational and Operational Characteristics

Predictive performance alone is insufficient to determine clinical or practical viability; computational efficiency and training dynamics are also critical.

ResNet18 was the most computationally efficient, requiring approximately two hours to complete 20 training epochs. Its relatively shallow depth and smaller parameter count make it well suited to environments with limited resources or real-time deployment constraints. DenseNet121, while delivering superior predictive metrics, incurred higher computational costs due to its dense connectivity pattern, but these were justified by its stability and robust convergence.

EfficientNet-B0, although theoretically optimised for parameter efficiency, required the longest training time in practice. In Google Colab, each epoch took nearly four hours, though this was reduced to around one hour per epoch on the Kaggle framework. The additional computational overhead arises from EfficientNet's compound scaling, which increases operational complexity. The Random Forest classifier had negligible training demands once feature embeddings were extracted, but its overall predictive capacity was limited by the quality of these features, highlighting its inferiority to end-to-end CNN pipelines.

4.3.3 Qualitative Interpretability of Predictions

To evaluate whether the models attended to clinically relevant structures, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to visualise salient regions influencing predictions.

All three CNNs consistently highlighted meaningful dermatoscopic features such as irregular pigmentation, atypical dots and globules, and lesion borders. This alignment with dermatological practice reinforces the clinical credibility of their predictions. Among the CNNs, DenseNet121 generated the most sharply localised and focused heatmaps, a likely benefit of its feature reuse mechanism, which preserves fine-grained spatial information. ResNet18 and EfficientNet-B0 produced broader heatmaps, but both remained aligned with the lesion regions of interest. EfficientNet-B0 also frequently extended attention to perilesional skin, suggesting that contextual cues were incorporated into its decision-making.

By contrast, the Random Forest model offers no native pixel-level interpretability, as it classifies based on extracted feature vectors rather than raw image data. This lack of visual explanation significantly reduces its suitability for clinical applications, where transparency and justification of decisions are essential.

4.3.4 Summary of Comparative Strengths

The comparative analysis highlights the distinct advantages of each model. EfficientNet-B0 demonstrated the highest sensitivity and F1-score, making it the most appropriate for applications where minimising false negatives is paramount. DenseNet121 achieved the highest precision and most reliable interpretability, rendering it the strongest candidate for clinical translation. ResNet18 proved to be the most computationally efficient, offering a balanced performance-to-resource ratio suitable for real-time or resource-constrained settings. The Random Forest classifier, while a useful baseline for benchmarking, was conclusively outperformed and is not suitable for clinical deployment in its current form.

Together, these findings demonstrate that CNNs not only provide superior predictive accuracy but also deliver clinically meaningful interpretability, establishing them as the preferred models for dermatological image classification. The broader implications of these comparative insights for clinical practice are explored in Chapter 5.

4.4 Application Output

The final stage of this project involved deploying the trained ResNet18 model as part of a prototype application, hosted on Hugging Face Spaces using the Streamlit framework. The aim of this deployment was to demonstrate how a high-performing convolutional neural network could be translated into a functional and accessible tool, while also highlighting issues of usability, transparency, and interpretability.

The application opens with a disclaimer page that informs the user the system is intended solely for research and educational purposes and should not be used as a substitute for professional medical advice. A QR code is also provided at this stage, which enables users to access the application directly on their mobile devices, offering the flexibility of capturing images with a phone camera in addition to uploading them from a computer.

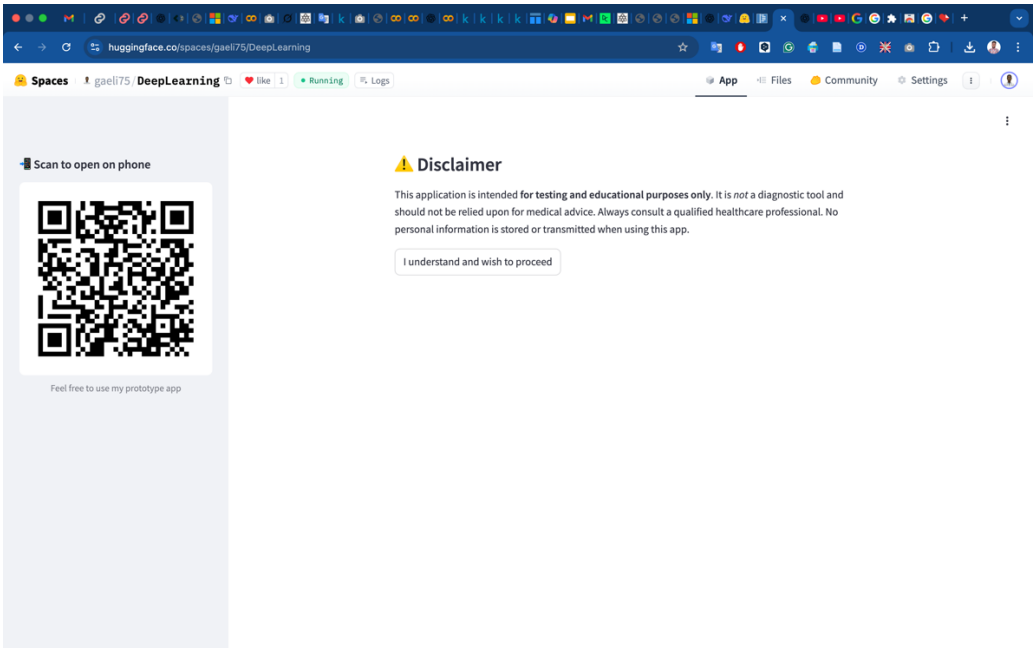


Figure 22.: Disclaimer page and QR code access to the application.

Once the disclaimer has been acknowledged, the user is taken to the main interface where images can either be uploaded or captured in real time. After the image is submitted, the model processes it through the same preprocessing pipeline employed during training, including resizing to 224×224 pixels and normalisation based on ImageNet mean and standard deviation values. The processed image is then passed through the fine-tuned ResNet18 network to generate predictions.

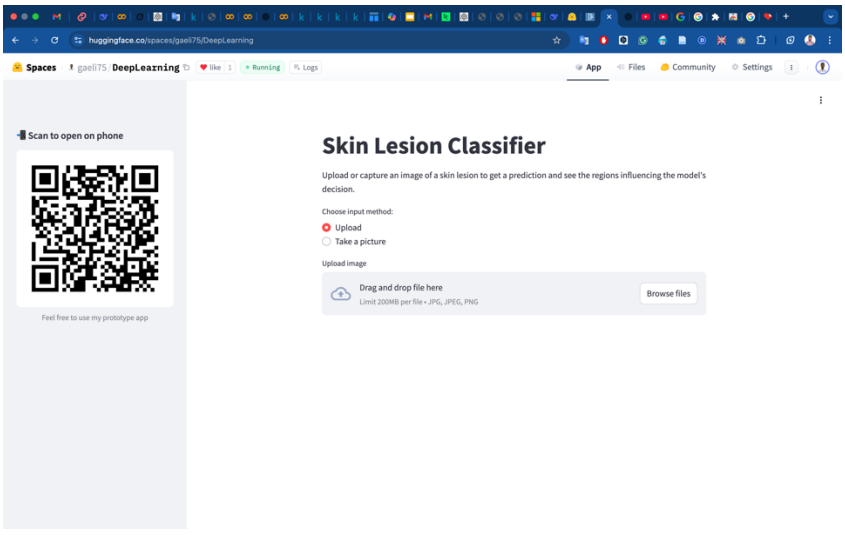


Figure 23.: Image input interface showing upload and capture options.

The output consists of the predicted lesion class, the associated confidence score, and a probability distribution across all seven diagnostic categories. These results are displayed alongside a Grad-CAM heatmap, which highlights the regions of the lesion most influential in the decision-making process. In addition, the application provides a short textual explanation of the predicted class and a corresponding recommendation. For example, if the model predicts melanoma, the system issues an urgent recommendation to seek medical attention, whereas in the case of a benign lesion such as a nevus, the message advises routine monitoring.

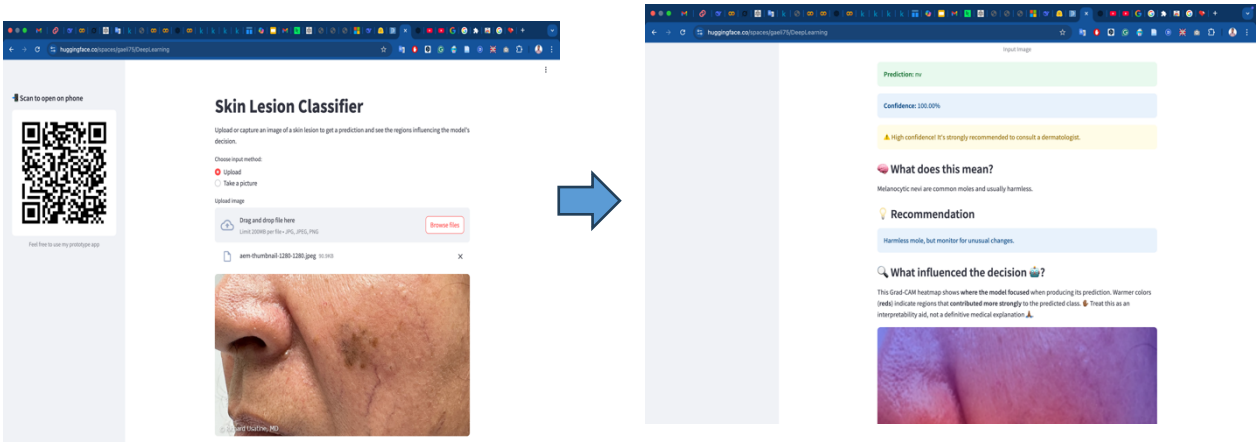


Figure 24.: Example of a user-uploaded dermatoscopic image.

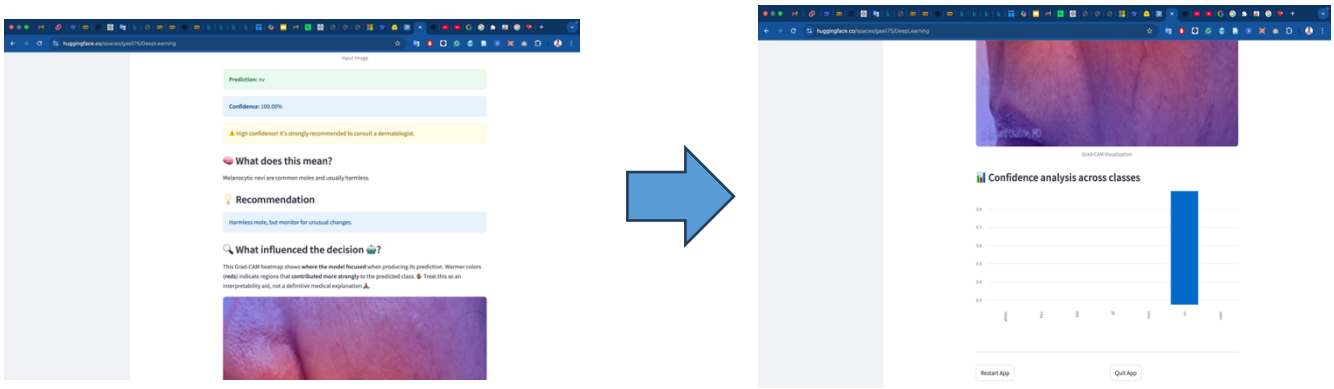


Figure 25.: Prediction output including class label, confidence score, and explanatory text.

The interface also includes a confidence distribution bar chart that visualises the likelihood of alternative classes. This feature encourages users to interpret the model's prediction not as a definitive outcome but as a distribution of probabilities, consistent with how diagnostic uncertainty is managed in clinical practice. Finally, the application provides restart and quit options so that the user may either test another image or terminate the session.

Figures 23 to 25 illustrate the complete workflow of the application, from the disclaimer and input screen through to the prediction results, Grad-CAM visualisations, and user options. Taken together, these outputs demonstrate that the deployed system successfully integrates classification, interpretability, and usability within a single interface. Although limited to research purposes, the prototype provides a proof of concept for how deep learning models for skin lesion classification can be embedded into accessible decision-support applications.

Chapter 5 – Discussion

This chapter provides a critical interpretation of the results presented in Chapter 4, situating the findings within the broader context of machine learning in medical imaging. It discusses the performance and trade-offs of the implemented models, the challenges of class imbalance, the critical insights gained from explainable AI (XAI) techniques, and the practical significance of the developed prototype application. The chapter concludes by honestly addressing the study's limitations and outlining the necessary steps for translating this research into a robust, clinically viable tool.

5.1 Interpretation of Model Performance

The quantitative evaluation demonstrated that all three deep convolutional neural networks ResNet18, DenseNet121, and EfficientNet-B0 achieved exceptional performance on the HAM10000 dataset, with all key metrics of accuracy, precision, recall, F1-score, exceeding 95%. This consistently high performance across architectures reaffirms the established literature on the suitability of deep learning for complex dermatoscopic image analysis [8, 49]. The success of these models is predicated on their ability to automatically learn hierarchical feature representations directly from pixel data, capturing nuances that are often challenging to quantify with handcrafted features.

The nuanced differences between the models, however, are highly informative. As summarized in Table 10, each architecture exhibited a distinct performance profile rooted in its fundamental design principles. ResNet18 established itself as a reliable and computationally efficient baseline. Its residual learning framework effectively mitigated the vanishing gradient problem, enabling stable training and solid performance [5]. This makes it an ideal choice for resource-constrained environments or for establishing a performance benchmark.

DenseNet121 achieved the highest precision, indicating its superior ability to minimize false positives. This is likely a direct benefit of its dense connectivity pattern, which promotes feature reuse throughout the network. This architecture allows later layers to access and refine low-level feature maps from earlier layers, leading to more discriminative and well-informed classifications [6, 39]. A model with high precision is particularly valuable in a clinical setting to prevent unnecessary patient anxiety and invasive procedures stemming from false alarms.

Conversely, EfficientNet-B0 achieved the highest recall and F1-score, indicating superior sensitivity and the best overall balance between precision and recall. Its compound scaling method, which uniformly optimizes network width, depth, and resolution, appears exceptionally effective at capturing the multi-scale features

necessary to identify malignant cases [7, 41]. A high-recall model is crucial for a screening tool, as its primary objective is to minimize false negatives a critical failure mode in oncology where a missed melanoma can have severe consequences [1].

The stark contrast with the Random Forest baseline, which struggled significantly, underscores a fundamental paradigm shift. It confirms that the performance leap in medical image analysis is not merely due to "more complex algorithms" but is intrinsically linked to the end-to-end feature learning capability of deep CNNs, which far surpasses the power of traditional models relying on pre-extracted features [3, 52].

Table 9: Comparative strengths of the evaluated models

Model	Primary Strength	Secondary Strength	Key Limitation
ResNet18	Computational Efficiency	Robust Baseline Performance	Slightly lower recall than more complex models
DenseNet121	Precision & Interpretability	Feature Reuse & Stable Convergence	Higher parameter count and memory usage
EfficientNet-B0	Recall & Overall F1-Score	Parameter Efficiency	Longest training time on limited hardware
Random Forest	Implementation Simplicity	N/A (Baseline)Poor performance;	Poor performance; no native pixel-level explainability

When contextualized within existing literature, the performance of these models is highly competitive. The seminal work by Esteva et al. [8] demonstrated dermatologist-level classification using a deep CNN trained on a very large (129,450-image) dataset. Brinker et al. [2] reported a CNN accuracy of 86.6% specifically for melanoma classification. The models in this study, achieving accuracies above 97% on the multi-class HAM10000 dataset with efficient architectures, highlight the effectiveness of modern transfer learning strategies. Furthermore, this work extends beyond the predictive metrics of earlier studies by integrating explainability methods, providing crucial insights into the models decision-making processes and strengthening the clinical relevance of the predictions.

5.2 Discussion on Class Imbalance Handling

Class imbalance is an inherent and well-documented challenge in medical image datasets, and the HAM10000 dataset is no exception [10, 48]. The heavy overrepresentation of melanocytic nevi (nv) compared to rarer but critical classes like

melanoma (mel) and basal cell carcinoma (bcc) poses a significant risk of biasing models towards the majority class.

The strategies employed including data augmentation by rotation, flipping and scaling. Stratified sampling were necessary and partially successful. They prevented the models from completely ignoring minority classes and ensured that each split of the data was representative of the overall distribution. However, the persistence of slightly lower recall metrics for the melanoma class, as noted in the confusion matrices, indicates that these standard techniques were insufficient to fully overcome the imbalance.

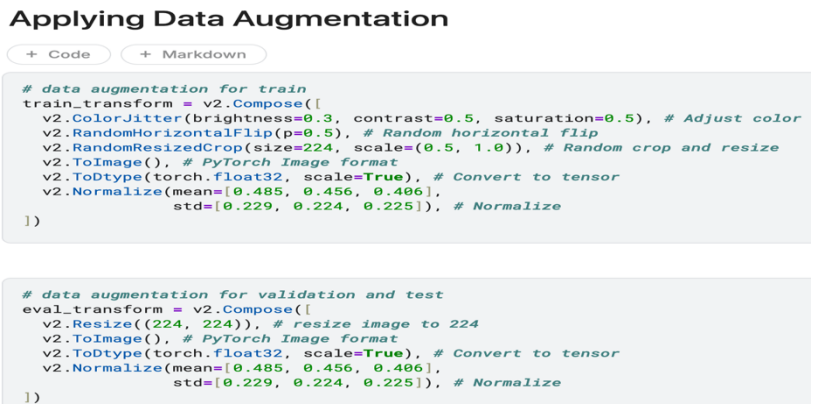


Figure 26: Applying Data augmentation

This residual bias highlights a crucial area for future work. While augmentation increases variability, it does not add new unique examples of minority classes. More advanced techniques are required to truly address this issue. These could include:

- Algorithmic-level approaches: Implementing cost-sensitive learning or using loss functions like Focal Loss [arXiv:1708.02002], which down-weights the loss assigned to well-classified examples, forcing the model to focus harder on difficult, minority-class cases.
- Data-level approaches: Employing synthetic data generation using Generative Adversarial Networks (GANs) to create realistic, high-quality synthetic dermatoscopic images of minority classes, thereby augmenting the dataset with truly new information [54].
- Hybrid approaches: Combining advanced oversampling techniques like SMOTE with traditional augmentation to create a more balanced and diverse training set.
- Addressing this imbalance is not merely an academic exercise; it is a prerequisite for developing models that are truly equitable and effective across all disease types.

5.3 Insights from Grad-CAM

Grad-CAM visualisations served as a critical tool for model validation and interpretation, confirming that the CNN models consistently attended to clinically meaningful regions within lesions, such as asymmetry, border irregularities, and atypical pigment patterns. This alignment with the ABCD rule of dermatoscopy [56] and the Chaos and Clues algorithm [57] provides strong evidence that the models learned pathologically relevant features.

Architectural differences yielded distinct interpretability profiles. DenseNet121 produced the sharpest and most localized heatmaps, a visual manifestation of its dense connectivity that preserves fine-grained spatial information [6]. ResNet18 and EfficientNet-B0 provided broader attention distributions, with EfficientNet-B0 often incorporating valuable contextual tissue information, consistent with its design [7]. Crucially, Grad-CAM also functioned as an error analysis tool, revealing occasional instances where model attention was diverted by artefacts like hair or image corners. This underscores the necessity of robust preprocessing to prevent models from latching onto spurious correlations.

This deep emphasis on explainability differentiates the present work from earlier studies that focused primarily on predictive performance. By integrating Grad-CAM throughout the analysis and into the deployed application, this study demonstrates a practical framework for embedding interpretability into the AI lifecycle, thereby making models more transparent, trustworthy, and aligned with clinical needs.

5.4 Usability and Significance of the Application

The development of the prototype web application represents a critical step in translating experimental research into a tangible tool with potential real-world impact. Deploying the ResNet18 model selected for its optimal balance of performance and efficiency demonstrated the practical feasibility of integrating a deep learning system into an accessible, user-friendly interface.

The application's design prioritizes transparency and decision-support. By presenting not just a binary prediction but also a confidence score, a probability distribution across all classes, and the Grad-CAM heatmap, the interface provides a holistic view of the model's reasoning. This empowers a potential user to understand the "why" behind the prediction and to weigh the AI's suggestion against their own clinical judgment. The inclusion of clear disclaimers is paramount, explicitly framing the tool as an educational prototype to aid in clinical decision-making and not to replace it.

This work highlights the often-underappreciated challenge of AI usability in medicine. A model's accuracy is necessary but not sufficient for adoption; the tool must also

integrate seamlessly into existing workflows, provide interpretable results, and manage user expectations responsibly. This prototype serves as a proof-of-concept that bridges the gap between algorithmic development and clinical application, providing a foundation for future work on user studies, clinical validation, and compliance with regulatory standards like those from the FDA(The U.S. Food and Drug Administration) or CE.

Compared to previous approaches, which often remained limited to laboratory settings, this study extended the contribution by hosting the model on Hugging Face Spaces with a user-friendly Streamlit interface. While studies such as Tschandl et al. [46] validated classification performance, they did not explore deployment in accessible formats for non-specialist users. The addition of mobile access via QR code and integrated visual explanations makes this prototype distinctive in bridging the gap between academic research and potential clinical usability.

5.5 Limitations

Despite the encouraging results, this study has several limitations that must be acknowledged to provide a balanced perspective and guide future research, as summarized in Table 11.

Table 10: Summary of key limitations and proposed mitigation strategies

Limitation	Impact on Study	Potential Mitigation for Future Work
Class imbalance	Reduced recall for underrepresented classes	Apply focal loss, oversampling, or GAN-based augmentation
Dataset bias	Limited generalisability across populations	Validate on multi-centre and multi-source datasets
Overfitting risk	Model may learn dataset-specific artefacts	Use stronger regularisation and k-fold cross-validation
Transfer learning	Sub-optimal features from natural images	Explore domain-specific pretraining (e.g., RadImageNet)
Computational limits	Restricted hyperparameter tuning and scaling	Employ distributed training on HPC or cloud platforms
Grad-CAM resolution	Limited granularity of interpretability	Complement with SHAP, LIME, or Integrated Gradients

First, the dataset bias inherent in HAM10000 must be considered. The images are sourced from a limited set of populations and clinical settings, which may not represent the full spectrum of skin types and disease presentations found globally. A model performing well on HAM10000 may not generalize well to populations with darker skin phototypes, where skin cancer often presents differently and is frequently diagnosed later.

Second, while measures were taken to prevent it, overfitting remains a persistent risk, especially for deeper, more complex models like DenseNet121 and EfficientNet-B0. Their high capacity allows them to potentially "memorize" subtle artefacts in the HAM10000 dataset rather than learning truly generalizable features of skin lesions.

Third, the reliance on transfer learning from ImageNet is a double-edged sword. While it is a proven and effective strategy, the domain shift between natural images (e.g., cats, cars) and medical dermatoscopic images is significant. The foundational features learned on ImageNet may not be optimal for capturing the specific textures and patterns relevant to dermatology.

Finally, while Grad-CAM is a powerful tool, it is an imperfect explainability method. It produces coarse heatmaps and provides a post-hoc explanation rather than revealing the model's true, causal reasoning process. The heatmaps can sometimes be misleading, and their interpretation requires careful consideration.

5.6 Generalisability to Clinical Settings

The ultimate test for any medical AI system is its performance in the messy, unpredictable environment of real-world clinical practice. The strong results on the curated HAM10000 dataset are a necessary first step, but they are not sufficient to guarantee clinical utility.

Achieving true generalisability requires external validation on large, multi-centre, and prospective datasets that encompass a wide range of imaging devices, patient demographics (age, sex, skin type), and clinical settings of primary care and specialist dermatology clinics. Furthermore, integration into clinical workflows presents its own set of challenges, including ensuring data privacy (e.g., HIPAA/GDPR compliance), meeting regulatory standards for software as a medical device (SaMD), and designing interfaces that save time rather than add to a clinician's cognitive load.

Nevertheless, the findings of this study are promising. They suggest that CNN-based models, particularly when augmented with interpretable outputs like Grad-CAM, have strong potential to function as clinical decision-support systems. In this capacity, they would not replace dermatologists but would act as a powerful second reader, helping to prioritize cases, reduce diagnostic errors, and improve overall efficiency. The path forward must involve close collaboration between AI researchers, clinicians, and regulators to design robust validation studies and translate these research prototypes into tools that are safe, effective, and equitable for all patients.

Unlike earlier studies that largely focused on controlled experimental conditions, this work demonstrated how a high-performing CNN model can be integrated into an interactive application, providing predictions, interpretability, and usability in a publicly accessible environment. While still limited to research purposes, this step

towards deployment highlights a practical contribution beyond accuracy metrics. The system illustrates how deep learning models can be embedded in decision-support applications, making them more approachable for clinicians and potentially accelerating translation into healthcare practice.

Chapter 6 – Future Work

6.1 Summary of Findings

This project has demonstrated the potential of deep learning methods for the automated classification of dermatoscopic images. By training and evaluating three convolutional neural networks, ResNet18, DenseNet121, and EfficientNet-B0 on the HAM10000 dataset, the study showed that accuracies above 97% can be consistently achieved. Each model displayed distinct strengths. ResNet18 proved to be a reliable and efficient baseline, DenseNet121 offered the highest precision, thereby reducing false positives and EfficientNet-B0 achieved the highest recall and F1-score, making it the most sensitive to true positive cases.

Beyond these standard measures, the models were also assessed using the area under the ROC curve which provides a more robust understanding of their discriminatory power across lesion categories. All three CNNs achieved macro-average AUC values above 0.95, confirming their ability to distinguish between malignant and benign classes across thresholds. These findings strengthen confidence in the models' reliability for potential screening use.

The Random Forest baseline was included as a traditional comparator. While useful in highlighting the advantage of handcrafted feature approaches, it underperformed substantially compared to CNNs and was not evaluated using AUC-ROC, since its outputs were not probabilistic in the same way as the deep models.

A central element of this work was the integration of Grad-CAM visualisations, which provided insight into model predictions and confirmed that the CNNs attended to clinically meaningful lesion regions. These interpretability outputs complemented the quantitative metrics, offering assurance that strong numerical results were grounded in clinically relevant image features. In addition, the deployment of the ResNet18 model as a prototype application on Hugging Face Spaces translated the research into a practical tool. The application offered predictions, confidence distributions, Grad-CAM heatmaps, and probability analyses in an accessible interface, illustrating how models can move beyond experimental evaluation towards user-oriented decision support.

6.2 Contributions of the Study

The contributions of this research are threefold. First, it provides empirical evidence that CNNs not only achieve high accuracy, precision, recall, and F1-scores, but also maintain strong performance under ROC-AUC evaluation, a metric widely regarded as one of the most reliable in medical AI. Second, it integrates explainability through Grad-CAM visualisations, allowing medical practitioners to inspect the reasoning

behind predictions and strengthening trust in automated outputs. Third, by deploying the ResNet18 model into a working application, the project demonstrates how research can transition towards practical tools suitable for web and mobile environments.

When contrasted with existing approaches, this study differentiates itself by combining strong numerical performance, including high ROC-AUC values for CNNs with interpretability and usability. Many earlier studies prioritised accuracy alone. However, om this work highlights that genuine clinical impact requires a broader balance of predictive performance, visual explanation, and accessible deployment.

6.3 Directions for Future Research

Despite promising results, there remain important areas for further investigation. Class imbalance within HAM10000 continued to affect the minority categories, particularly melanoma, where recall, F1, and AUC-ROC values were slightly lower than for benign lesions. Future work should explore more advanced strategies for imbalance handling, such as focal loss, generative augmentation, or the integration of cost-sensitive learning. Broader validation is also needed. While HAM10000 is a benchmark dataset, it is limited in demographic diversity and clinical context. Evaluation across multi-centre and multi-source datasets would provide stronger evidence of generalisability.

Methodologically, there is scope to expand beyond CNNs by exploring hybrid architectures that incorporate transformer-based models or attention mechanisms. Ensemble learning approaches could also be considered, combining the complementary strengths of models such as ResNet and EfficientNet to improve robustness. In terms of evaluation, future research should consistently report ROC-AUC alongside accuracy-based metrics, since AUC captures class separability and is more resilient to imbalance.

Finally, future research should extend the prototype application towards clinical validation. This would involve usability testing with dermatologists, integration with teledermatology platforms, and alignment with regulatory and ethical frameworks. The deployment on Hugging Face Spaces provides a strong proof of concept, but further steps are necessary to move from research to real-world impact.

Chapter 7 – Conclusion

This dissertation set out to explore the potential of deep learning models for the automated classification of dermatoscopic images, with the overarching goal of improving early detection of skin cancer. Through a comparative study of ResNet18, DenseNet121, and EfficientNet-B0, supported by a Random Forest baseline, it has been shown that convolutional neural networks can achieve accuracies above 97% on the HAM10000 dataset, while also delivering macro-average ROC-AUC values above 0.95. These results confirm not only that CNNs are capable of achieving high raw accuracy, but also that they possess strong discriminatory power across thresholds, a property essential for medical screening where both sensitivity and specificity must be balanced.

The research has demonstrated that technical performance alone is not sufficient for clinical adoption. By incorporating interpretability through Grad-CAM and translating the ResNet18 model into a working prototype deployed on Hugging Face Spaces, the project moved beyond experimental analysis to emphasise usability, transparency, and accessibility. This integration of predictive performance, interpretability, and deployment constitutes the central contribution of the study.

At the same time, the work recognises its limitations. Dataset imbalance and limited diversity constrained generalisability, and computational restrictions limited exploration of alternative architectures and hyperparameter configurations. In addition, while Grad-CAM offered meaningful interpretability, it cannot fully resolve all ambiguities in model reasoning. These limitations highlight the importance of continued research into dataset diversity, explainability techniques, and clinical validation.

Looking forward, the findings suggest that lightweight but interpretable deep learning models, validated through both accuracy-based and ROC-AUC metrics, could play a significant role in augmenting dermatological practice. By supporting clinicians with reliable predictions and visual explanations, AI-based decision-support systems have the potential to enhance diagnostic accuracy and widen access to early detection, particularly in resource-constrained healthcare environments. This study provides a step in that direction, offering both a demonstration of technical feasibility and a proof of concept for practical deployment.

References

[1] R. L. Siegel, K. D. Miller, and N. S. Wagle, “Cancer statistics, 2023,” *CA: A Cancer J. Clin.*, vol. 73, no. 1, pp. 17–48, 2023.

[2] T. J. Brinker et al., “Deep neural networks are superior to dermatologists in melanoma image classification,” *Eur. J. Cancer*, vol. 119, pp. 11–17, 2019.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[7] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[8] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[10] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci. Data*, vol. 5, 180161, 2018.

[11] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020.

[12] American Cancer Society, “Cancer Facts & Figures 2023,” Atlanta: American Cancer Society, 2023.

[13] A. Piccolo, A. Ferrari, R. Peris, and G. Argenziano, “Reproducibility of dermoscopic diagnosis of melanoma and melanocytic nevi,” *Arch. Dermatol.*, vol. 140, no. 5, pp. 579–584, 2004.

[14] M. Karimkhani et al., “Global burden of skin disease as reflected in Cochrane Database of Systematic Reviews,” *JAMA Dermatol.*, vol. 150, no. 9, pp. 945–951, 2014.

[15] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset: A large

collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, 180161, 2018.
[16] G. Xie, H. Zhang, and X. Shen, "Hair and ruler mark removal in dermoscopy images using deep learning," *Comput. Med. Imaging Graph.*, vol. 86, p. 101802, 2020.
[17] S. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *IEEE Syst. J.*, vol. 8, no. 3, pp. 965–979, 2014.
[18] L. Combalia and N. Codella, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv preprint arXiv:1902.03368*, 2019.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[20] S. Armato III et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011.

[21] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. CVPR*, 2017, pp. 2097–2106.

[22] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, 2016, pp. 2818–2826.

[23] N. Combalia et al., "BCN20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[25] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[26] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.

[27] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. NeurIPS*, 2017, pp. 5574–5584.

[28] A. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Proc. NeurIPS*, 2019, pp. 3342–3352.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[30] Z. Chen, X. Fan, J. Jin, and J. Xu, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[31] J. Long et al., "Transfer learning for medical image classification," *Med. Image Anal.*, vol. 42, pp. 60–73, 2017.

[32] Y. Liu et al., “On-device dermatology AI,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 3, pp. 497–507, 2022.
[33] M. T. Ribeiro et al., “Visual texture analysis in dermoscopy,” *J. Med. Imaging*, vol. 8, no. 4, p. 044502, 2021.
[34] C. Barata et al., “Impact of resolution on melanoma diagnosis,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 5, pp. 2149–2160, 2022.
[35] L. R. Arantes et al., “Depth vs accuracy in skin cancer models,” *Comput. Biol. Med.*, vol. 153, p. 106519, 2023.
[36] S. Qiao et al., “Adaptive sampling for dermoscopy,” *Med. Image Anal.*, vol. 78, p. 102431, 2022.
[37] P. Tschandl et al., “Attention mechanisms in dermatology AI,” *Nat. Mach. Intell.*, vol. 4, no. 11, pp. 927–938, 2022.
[38] Z. Ge et al., “Dilated CNNs for lesion analysis,” *IEEE Trans. Med. Imaging*, vol. 41, no. 4, pp. 881–892, 2022.

[39] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.
[40] A. Tajbakhsh et al., “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.
[41] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[42] N. Y. Tang et al., “A lightweight EfficientNet model for skin lesion classification using dermoscopy images,” *Comput. Biol. Med.*, vol. 133, p. 104402, 2021.
[43] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, and K. Maier-Hein, “nnU-Net: Self-adapting framework for U-Net-based medical image segmentation,” *Nat. Methods*, vol. 18, pp. 203–211, 2021.

[44] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020.
[45] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[46] P. Tschandl, C. Rosendahl, and H. Kittler, “Human–computer collaboration for skin cancer recognition,” *Nat. Med.*, vol. 26, no. 8, pp. 1229–1234, 2020.
[47] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. NeurIPS*, 2017, pp. 4765–4774.

[48] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci. Data*, vol. 5, 180161, 2018.
[49] T. J. Brinker et al., “Convolutional neural networks are superior to dermatologists in melanoma image classification,” *Eur. J. Cancer*, vol. 119, pp. 11–17, 2019.
[50] A. Mahbod, R. Schaefer, G. Wang, and H. Eghbalnia, “Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification,” *Comput. Methods Programs Biomed.*, vol. 197, p. 105765, 2020.
[51] B. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[52] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques, “Two systems for the detection of melanomas in dermoscopy images using texture and color features,” *IEEE Syst. J.*, vol. 8, no. 3, pp. 965–979, 2014.
[53] J. Kawahara and G. Hamarneh, “Multi-resolution-tract CNN with hybrid pre-training for skin lesion classification,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 164–172.
[54] N. Combalia et al., “BCN20000: Dermoscopic lesions in the wild,” *arXiv preprint arXiv:1908.02288*, 2019.
[55] D. M. W. Powers, “Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
[56] W. Stolz, O. Braun-Falco, P. Bilek, M. Landthaler, W. H. C. Burgdorf, and A. B. Cognetta, *Color Atlas of Dermatoscopy*. Oxford: Blackwell Science, 1994, pp. 18-23.
[57] C. Rosendahl, G. Cameron, I. H. McColl, and D. Wilkinson, "Dermatoscopy in routine practice: 'Chaos and Clues'," *Aust. Fam. Physician*, vol. 41, no. 7, pp. 482–487, Jul. 2012.
[58] Y. Haibo, Z. Liu, and F. Wang, “Evaluation metrics in medical imaging deep learning: A comprehensive review,” *Medical Image Analysis*, vol. 77, p. 102357, 2022.

Appendix

Appendix A - Terms Of Reference

Department of Computing and Mathematics Computing and Digital Technology Postgraduate Programmes Terms of Reference Coversheet	
Student name:	Fundo Lazaro Ambrosio Fernando
University I.D.:	24843141
Academic supervisor:	Dr Philip Sinclair
External collaborator (optional):	
Project title:	Using ResNet18 to Classify Skin Lesions from Medical Images: A Statistical Learning Approach
Degree title:	MSc Data Science
Project unit code:	6G7ZV0007
Credit rating:	60
Start date:	20 May 2025
ToR date:	7 May 2025
Intended submission date:	29 August 2025
Signature and date student:	
Signature and date external collaborator (if involved):	

1. Introduction

Skin cancer is among the most common cancers globally, with melanoma and non-melanoma types contributing significantly to the global disease burden. Early detection of malignant lesions is critical for improving treatment outcomes and patient survival [7]. Traditional diagnostic methods, such as visual inspection and dermoscopy, are time-consuming and rely heavily on the expertise of dermatologists, often leading to variability in diagnostic accuracy [8].

Recent advances in machine learning, particularly deep learning, have shown strong potential in automating skin lesion analysis, thereby supporting clinical decision-making. Convolutional Neural Networks (CNNs) have demonstrated performance comparable to that of dermatologists in classifying dermoscopic images [9][10]. Among these, ResNet architectures have been widely adopted due to their ability to mitigate the vanishing gradient problem and enable deeper networks [11]. This project focuses on the development and evaluation of ResNet18 for skin lesion classification and further compares its performance with other state-of-the-art CNN architectures such as DenseNet [12] and EfficientNet [13], both of which have shown competitive results in recent medical imaging studies.

In addition to model development, a digital interface (web or mobile application) will be created to visualize and interpret predictions, promoting better user interaction and clinical usability. This approach aligns with the broader goal of enhancing model accuracy, interpretability, and real-world applicability in healthcare settings [14].

2. Aim

To build and evaluate a robust deep learning-based classification model using ResNet18 [1] and other comparative architectures for dermoscopic image analysis, and to develop a companion application for result visualization and interpretation.

3. Objectives

- Investigate and implement ResNet18 [1] for classifying dermoscopic skin lesion images.
- Experiment with additional models (e.g., DenseNet121[2], EfficientNet-B0 [3]) to compare performance, focusing on improving precision.
- Evaluate all models using classification metrics including accuracy, precision, recall, and F1-score.
- Use Grad-CAM [5] to provide visual explanation of model decisions.
- Develop a mobile or web application for interfacing with the models and visualizing predictions.
- Analyse comparative results and highlight insights regarding model performance on minority classes.

4. Scope

This project is limited to the use of publicly available datasets, primarily the HAM10000 dataset of dermoscopic images[4]. The research focuses on the classification of these images into predefined lesion categories using CNN-based models. The project does not aim for clinical deployment but serves as a proof-of-concept for diagnostic support tools. While ResNet18 is the primary model, other architectures will be considered for comparative evaluation. A prototype app will be developed to enhance user interaction with model outputs.

5. Methodology

Dataset Acquisition and Preprocessing: The HAM10000 dataset will be used from Kaggle Datasets[4]. I plan to clean data, augmented (random flips, color jittering), normalized, and split into training/validation sets with an 80/20 ratio.

A structured literature review will be conducted to assess the adequacy, performance, and applicability of CNN models like ResNet18, DenseNet, and EfficientNet in dermatological image classification tasks. This review will help contextualize model choices and evaluation metrics.

- Model Development:
  - The starter point is to fine-tune a pretrained ResNet18 model for lesion classification.[1]
  - Research and test out other models such as DenseNet121 and EfficientNet to determine improvements in precision.[2][3]
- Training Strategy:
  - Use transfer learning and fine-tuning.
  - Cross-entropy loss and Adam optimizer with weight decay.[6]
  - Model evaluation based on classification metrics and confusion matrix.
- Visual Interpretability:
  - Apply Grad-CAM to interpret and visualize model focus on lesion areas.[5]
- Application Development:
  - Build a simple mobile or web app (using tools like Streamlit or Flutter).
  - App will display predictions, model comparisons, and Grad-CAM visualizations.
- Analysis:
  - Compare models on precision and recall, especially for minority lesion classes.
  - Document findings and critical evaluation in final report.

6. Deliverables

- Cleaned and augmented HAM10000 dataset.
- Trained ResNet18 and experimental CNN models.
- Comparative analysis and evaluation metrics.
- Grad-CAM visualizations and confusion matrices.
- Mobile or web-based application prototype.
- Final project dissertation in IEEE format.

7. Timeline

Phase	Dates
Literature Review	1 – 10 June 2025
Data Preparation	10 – 15 June 2025
Model Development (ResNet18)	16 – 25 June 2025
Alternative Models	26 June – 10 July 2025
App Development	11 – 20 July 2025
Evaluation & Visuals	21 – 31 July 2025
Report Writing	1 – 20 August 2025
Final Review and Submission	21 – 29 August 2025

8. Learning Outcomes

- Apply statistical learning and deep learning models to solve real-world classification problems in healthcare.
- Understand and compare the interpretability, precision, and adequacy of various CNN architectures including ResNet18 [1].
- Use visual tools such as Grad-CAM [5] to communicate model behavior and decision rationale .
- Develop a functional prototype application (web/mobile) to support the visualization and interaction with deep learning model outputs.
- Critically evaluate the performance and usability of AI models for medical applications.
- Critically evaluate and synthesize existing research on CNN architectures, particularly ResNet18, applied to skin lesion classification.

9. Roles and Responsibilities

I, Fundo Fernando, I am solely responsible for all aspects of the project including dataset handling, model training, performance evaluation, app development, and report writing. Additionally, ensuring all data ethics, integrity of code, and timely submission.

10. References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770-778, 2016. [Online]. Available: https://arxiv.org/abs/1512.03385

[2] G. Huang et al., "Densely connected convolutional networks," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4700–4708, 2017.

[3] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proc. ICML*, pp. 6105–6114, 2019.

[4] F. Kabir Samanta, "Skin Cancer Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/farjanakabirsamanta/skin-cancer-dataset

[5] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013. [Online]. Available: https://www.statlearning.com/

[7] American Cancer Society. (2024). Cancer Facts & Figures 2024. American Cancer Society.
Josep Malvehy, Oleg Tschandl, Harald Kittler. (2020).

[8] The impact of clinical context in dermoscopic diagnosis: a web-based survey. Journal of the European Academy of Dermatology and Venereology. 34(3), pp. 592–597.

[9] 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, pp. 115–118.

[10] Philipp Tschandl, Cliff Rosendahl, Harald Kittler. (2020). Human–computer collaboration for skin cancer recognition. Nature Medicine. 26, pp. 1229–1234.
[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

[12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. (2017). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269.

[13] Mingxing Tan and Quoc V. Le. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the International Conference on Machine Learning (ICML), pp. 6105–6114.
[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.

Appendix B - Code Listings

This appendix presents selected code excerpts that were central to the implementation of the deep learning models, baseline experiments, and application prototype developed in this study. The full source code and notebooks are available in the project repository. Only essential blocks are included here for clarity and reproducibility.

Listing B.1: ResNet18 model definition

```
# get a pretrain resnet18
model = torchvision.models.resnet18(weights='IMAGENET1K_V1')
```

```
Downloading: "https://download.pytorch.org/models/resnet18-f37072fd.pth" to /root/.cache/torch/hub/checkpoints/resnet18-f37072fd.pth
100%|██████████| 44.7M/44.7M [00:00<00:00, 171MB/s]
```

```
# Add a new layer/change the last layer
model.fc = nn.Linear(model.fc.in_features, num_class)
model.to(device)
```

Listing B.3: DenseNet121 model definition

Model

```
from torchvision.models import densenet121
```

```
# geting a pretrain resnet18
model = densenet121(pretrained=True)
```

```
/usr/local/lib/python3.12/dist-packages/torchvision/models/_utils.py:208: UserWarning: The parameter 'pretrained' is deprecated since 0.13 and may be removed in the future, please use 'wei
  warnings.warn(
/usr/local/lib/python3.12/dist-packages/torchvision/models/_utils.py:223: UserWarning: Arguments other than a weight enum or `None` for 'weights' are deprecated since 0.13 and may be remov
  warnings.warn(msg)
Downloading: "https://download.pytorch.org/models/densenet121-a639ec97.pth" to /root/.cache/torch/hub/checkpoints/densenet121-a639ec97.pth
100%|██████████| 30.8M/30.8M [00:00<00:00, 152MB/s]
```

```
# Add a new layer/ to change the last layer
num_classes = 7
model.classifier = nn.Sequential(
    nn.Linear(model.classifier.in_features, num_classes)
)
model.to(device)
```

Listing B.4: EfficientNet-B0 model definition

```
# get a pretrain EfficientNet-B0


# Load pretrained EfficientNet-B0
model = models.efficientnet_b0(weights=EfficientNet_B0_Weights.IMAGENET1K_V1)
```

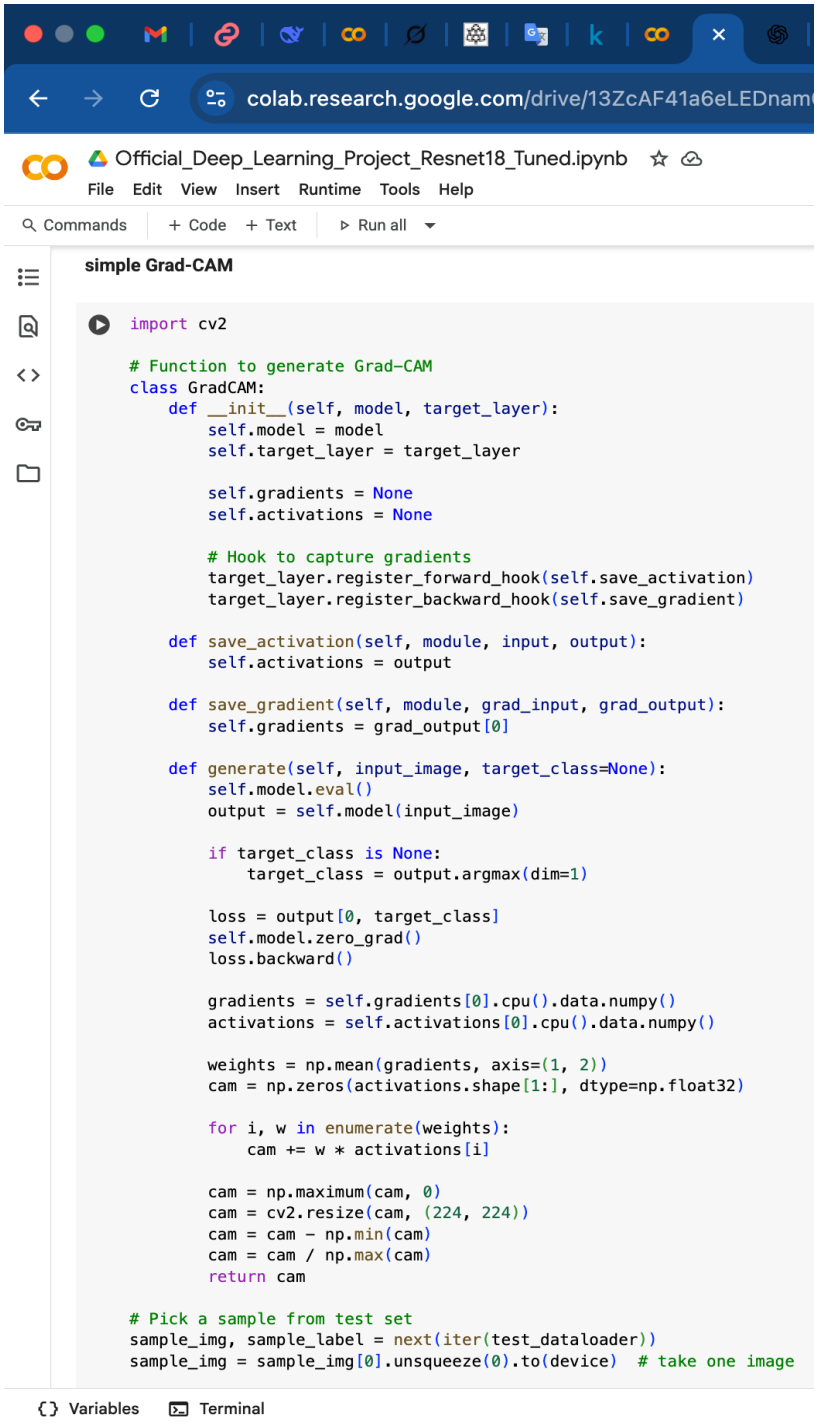
```
Downloading: "https://download.pytorch.org/models/efficientnet_b0_rwightman-7f5810bc.pth" to /root/.cache/torch/hub/checkpoints/efficientnet_b0_rwightman-
7f5810bc.pth
100%|██████████| 20.5M/20.5M [00:00<00:00, 142MB/s]
```

+ Code + Markdown

```
in_features = model.classifier[1].in_features  # 1280 for B0
model.classifier[1] = nn.Linear(in_features, num_class)

model = model.to(device)
```

Listing B.5: Grad-CAM function implementation

The Grad-CAM function used forward and backward hooks to capture activations and gradients from the final convolutional block, producing heatmaps that were overlaid on input images.


colab.research.google.com/drive/13ZcAF41a6eLEDnam
Official_Deep_Learning_Project_Resnet18_Tuned.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
simple Grad-CAM
import cv2
# Function to generate Grad-CAM
class GradCAM:
def __init__(self, model, target_layer):
self.model = model
self.target_layer = target_layer
self.gradients = None
self.activations = None
# Hook to capture gradients
target_layer.register_forward_hook(self.save_activation)
target_layer.register_backward_hook(self.save_gradient)
def save_activation(self, module, input, output):
self.activations = output
def save_gradient(self, module, grad_input, grad_output):
self.gradients = grad_output[0]
def generate(self, input_image, target_class=None):
self.model.eval()
output = self.model(input_image)
if target_class is None:
target_class = output.argmax(dim=1)
loss = output[0, target_class]
self.model.zero_grad()
loss.backward()
gradients = self.gradients[0].cpu().data.numpy()
activations = self.activations[0].cpu().data.numpy()
weights = np.mean(gradients, axis=(1, 2))
cam = np.zeros(activations.shape[1:], dtype=np.float32)
for i, w in enumerate(weights):
cam += w * activations[i]
cam = np.maximum(cam, 0)
cam = cv2.resize(cam, (224, 224))
cam = cam - np.min(cam)
cam = cam / np.max(cam)
return cam
Pick a sample from test set
sample_img, sample_label = next(iter(test_dataloader))
sample_img = sample_img[0].unsqueeze(0).to(device) # take one image
Variables Terminal

colab.research.google.com/drive/13ZcAF41a6eLEDnamOT_kv

Official_Deep_Learning_Project_Resnet18_Tuned.ipynb

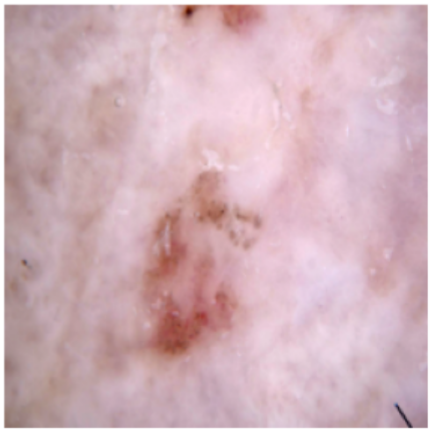
File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all

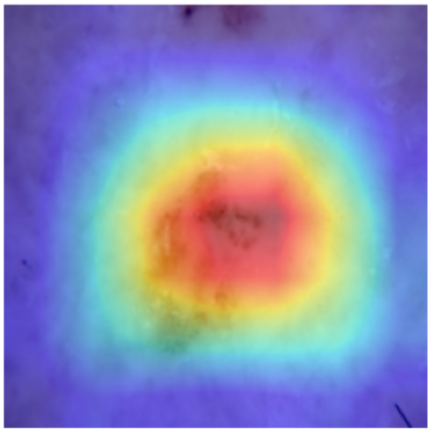
```
# Apply Grad-CAM
target_layer = model.layer4[-1]  # usually last layer for ResNet18
gradcam = Gra el, target_layer)
cam = gradcam. e(sample_img)

# Plot original + heatmap
img = sample_img.cpu().squeeze().permute(1,2,0).numpy()
img = (img - img.min()) / (img.max() - img.min())  # normalize image

plt.figure(figsize=(8,4))
plt.subplot(1,2,1)
plt.title("Original Image")
plt.imshow(img)
plt.axis('off')

plt.subplot(1,2,2)
plt.title("Grad-CAM Heatmap")
plt.imshow(img)
plt.imshow(cam, cmap='jet', alpha=0.5)  # overlay Grad-CAM
plt.axis('off')
plt.show()
```

```
/usr/local/lib/python3.11/dist-packages/torch/nn/modules/module.py:1830: Futu
  self._maybe_warn_non_full_backward_hook(args, result, grad_fn)
```

Original Image

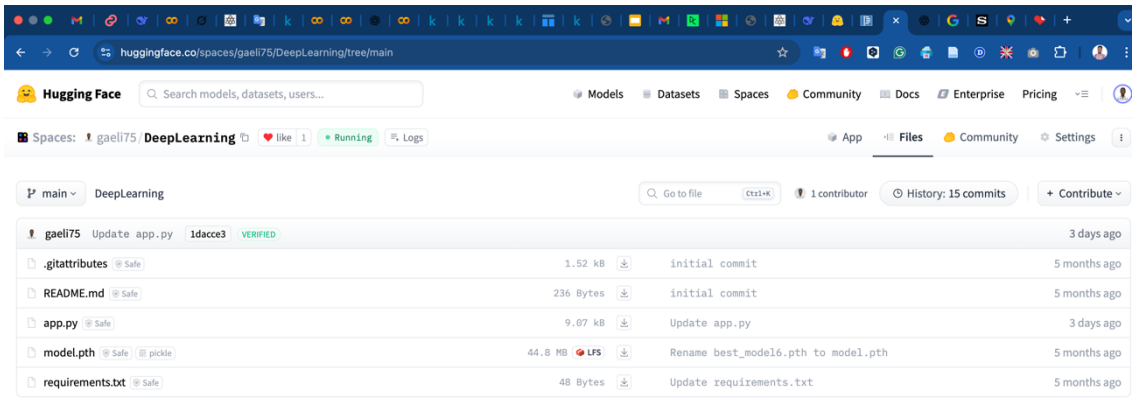


Grad-CAM Heatmap



Listing B.6: Application Deployment Streamlit Prototype

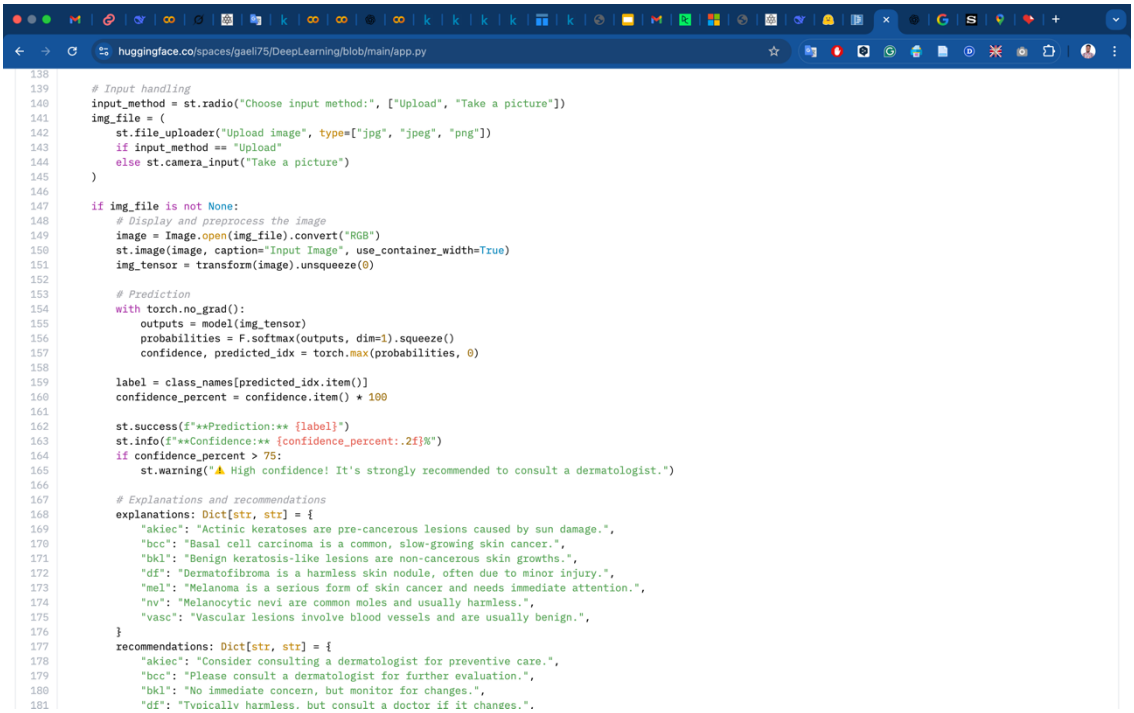
The trained ResNet18 model was deployed using Streamlit and hosted on Hugging Face Spaces. The app included a disclaimer page, image input options, prediction output with confidence values, Grad-CAM visualisation, and a probability distribution chart.



Inside the App.py file - model loading

```python
@st.cache_resource
def load_model() -> nn.Module:
    """Load the trained model from disk and prepare it for inference."""
    model = models.resnet18(pretrained=False)
    model.fc = nn.Linear(model.fc.in_features, len(class_names))
    model.load_state_dict(torch.load("model.pth", map_location=torch.device("cpu")))
    model.eval()
    return model
```

Listing B.7: Streamlit interface

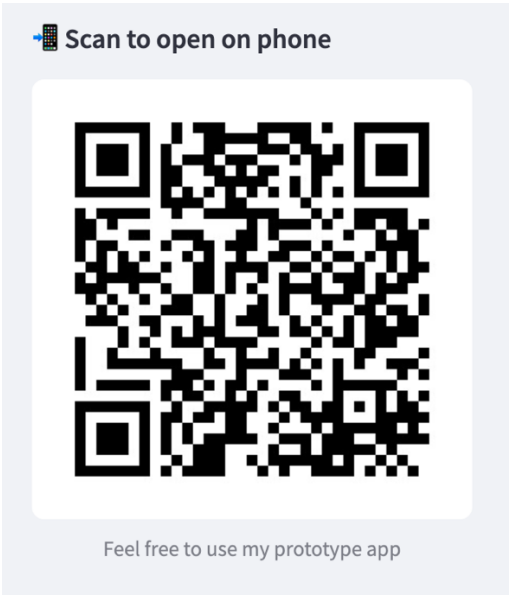


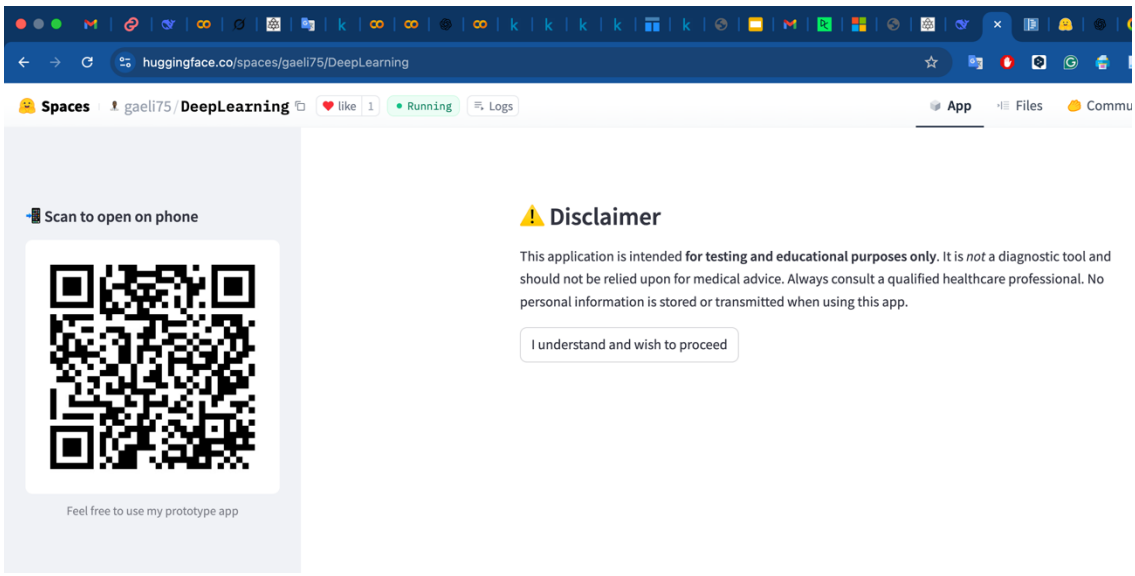

```python

    # Input handling
    input_method = st.radio("Choose input method:", ["Upload", "Take a picture"])
    img_file = (
        st.file_uploader("Upload image", type=["jpg", "jpeg", "png"])
        if input_method == "Upload"
        else st.camera_input("Take a picture")
    )

    if img_file is not None:
        # Display and preprocess the image
        image = Image.open(img_file).convert("RGB")
        st.image(image, caption="Input Image", use_container_width=True)
        img_tensor = transform(image).unsqueeze(0)

        # Prediction
        with torch.no_grad():
            outputs = model(img_tensor)
            probabilities = F.softmax(outputs, dim=1).squeeze()
            confidence, predicted_idx = torch.max(probabilities, 0)

        label = class_names[predicted_idx.item()]
        confidence_percent = confidence.item() * 100

        st.success(f"**Prediction:** {label}")
        st.info(f"**Confidence:** {confidence_percent:.2f}%")
        if confidence_percent > 75:
            st.warning("⚠️ High confidence! It's strongly recommended to consult a dermatologist.")

        # Explanations and recommendations
        explanations: Dict[str, str] = {
            "akiec": "Actinic keratoses are pre-cancerous lesions caused by sun damage.",
            "bcc": "Basal cell carcinoma is a common, slow-growing skin cancer.",
            "bkl": "Benign keratosis-like lesions are non-cancerous skin growths.",
            "df": "Dermatofibroma is a harmless skin nodule, often due to minor injury.",
            "mel": "Melanoma is a serious form of skin cancer and needs immediate attention.",
            "nv": "Melanocytic nevi are common moles and usually harmless.",
            "vasc": "Vascular lesions involve blood vessels and are usually benign.",
        }
        recommendations: Dict[str, str] = {
            "akiec": "Consider consulting a dermatologist for preventive care.",
            "bcc": "Please consult a dermatologist for further evaluation.",
            "bkl": "No immediate concern, but monitor for changes.",
            "df": "Typically harmless, but consult a doctor if it changes.",
```

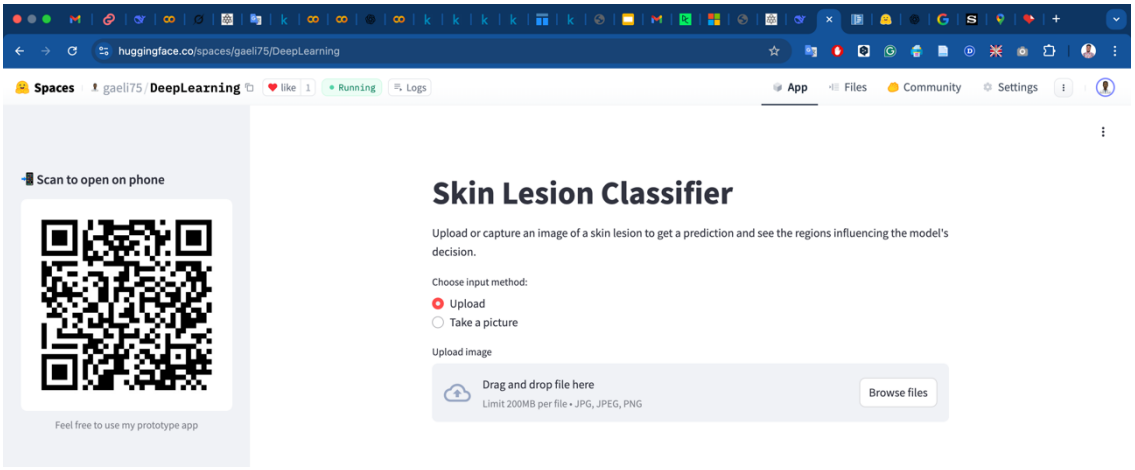
Appendix D – Application Screeshots

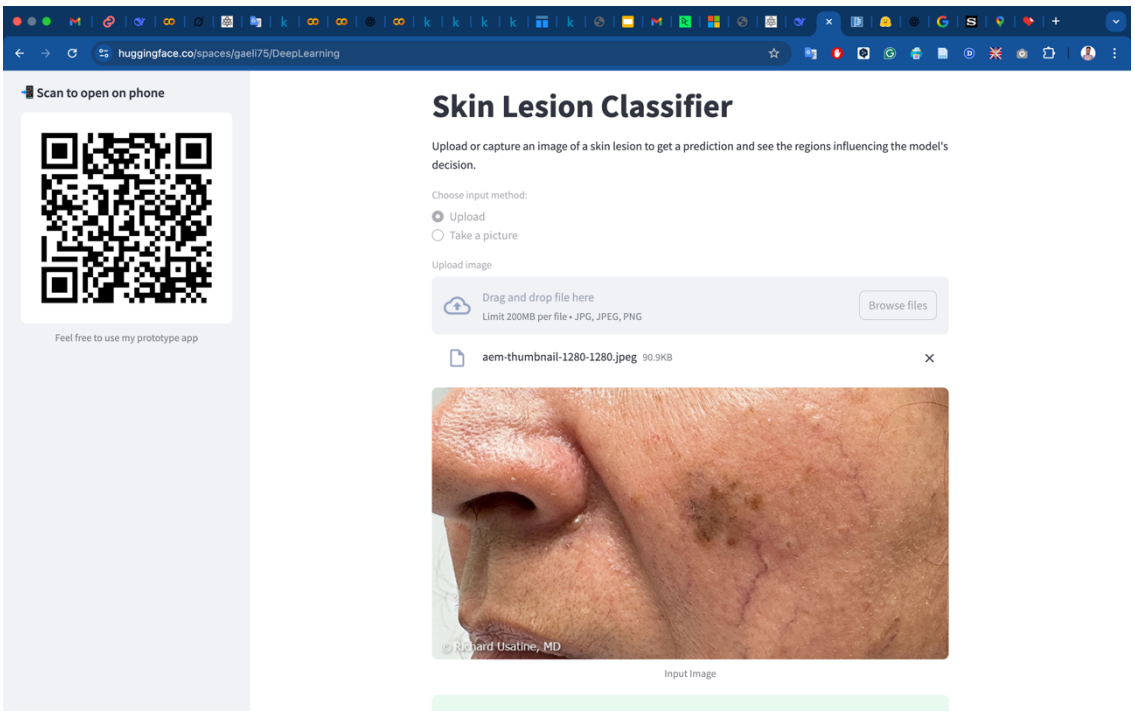
This appendix presents screenshots of the deployed application, which was implemented using the Streamlit framework and hosted on Hugging Face Spaces. The app provides an accessible interface for testing the ResNet18 skin lesion classifier, offering predictions, confidence values, Grad-CAM visualisations, and class probability distributions.
To facilitate access, the application is available online at the following link: Here

For convenience, a QR code has also been generated to allow users to access the application directly from a mobile device.




huggingface.co/spaces/gaeli75/DeepLearning
Spaces
gaeli75 / DeepLearning
like 1
Running
Logs
App
Files
Commu
Scan to open on phone
Feel free to use my prototype app
⚠️ Disclaimer
This application is intended for testing and educational purposes only. It is not a diagnostic tool and should not be relied upon for medical advice. Always consult a qualified healthcare professional. No personal information is stored or transmitted when using this app.
I understand and wish to proceed


huggingface.co/spaces/gaeli75/DeepLearning
Spaces
gaeli75 / DeepLearning
like 1
Running
Logs
App
Files
Community
Settings
Scan to open on phone
Feel free to use my prototype app
Skin Lesion Classifier
Upload or capture an image of a skin lesion to get a prediction and see the regions influencing the model's decision.
Choose input method:
Upload
Take a picture
Upload image
Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG
Browse files



Prediction: nv

Confidence: 100.00%

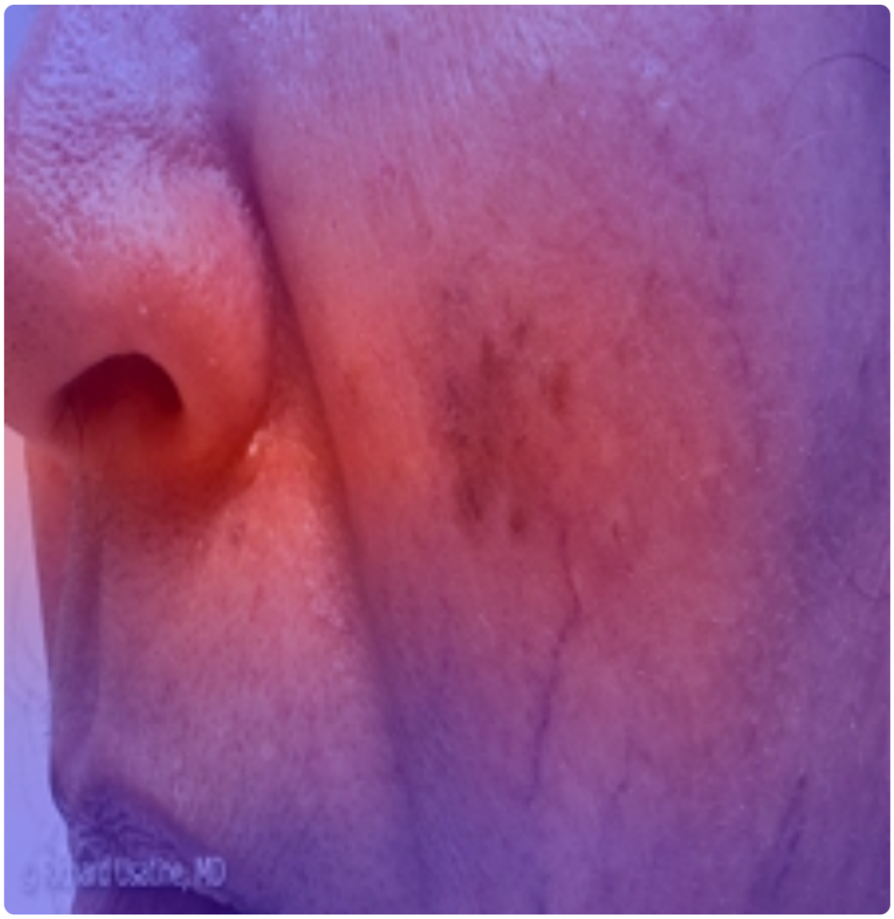
⚠️ High confidence! It's strongly recommended to consult a dermatologist.

🧠 What does this mean?

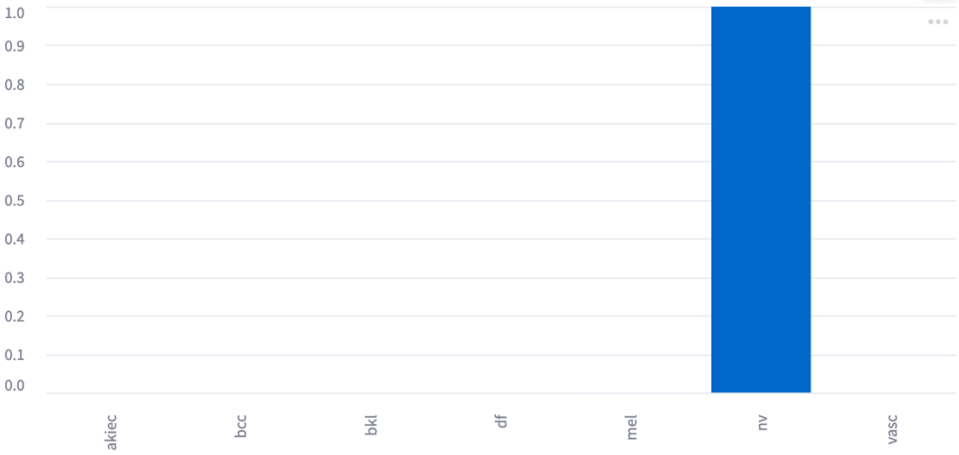
Melanocytic nevi are common moles and usually harmless.

🔍 What influenced the decision 🤖?

This Grad-CAM heatmap shows **where the model focused** when producing its prediction. Warmer colors (**reds**) indicate regions that **contributed more strongly** to the predicted class. 🖐🏾 Treat this as an interpretability aid, not a definitive medical explanation 🙏🏾.



Confidence analysis across classes



Restart App

Quit App