AI Enabled Skin Cancer Detection

Thesis Research

By Panagioths - Dimitrios Nikolakopoulos

May 2024

Contents Page

Project summary

The focus of this thesis is the development of AI models for skin cancer detection using machine learning and deep learning techniques. The models were trained on a large collection of dermoscopic images to improve early-stage skin cancer diagnosis. Collaborating closely with medical experts in dermatology and oncology, the research aims to provide a solution for rapid and accurate skin lesion classification, which can be particularly beneficial for remote areas with limited access to healthcare.

The primary objective was to explore the efficacy of different AI models—namely Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Xception—in classifying various types of skin lesions. A CNN-based model achieved a high accuracy of 98%, outperforming other architectures due to its efficiency with limited data. However, ViTs showed promise as a state-of-the-art alternative with different advantages, particularly in global feature extraction. Xception, another advanced architecture, provided additional insights into feature representation and model performance.

The data was sourced from the HAM10000 dataset, which contains over 10,000 labeled images of seven distinct skin lesion types. To address class imbalance and improve model performance, various data augmentation techniques were employed. The project also involved careful preprocessing, such as resizing and normalizing images to align with model requirements.

Throughout the research, a structured methodology was adopted, starting with exploratory data analysis to understand lesion distribution by demographic factors like age and gender. Subsequent cycles of model development involved training, refining, and evaluating multiple models to achieve optimal accuracy and generalizability. Each model's performance was critically assessed using metrics such as accuracy, precision, recall, and confusion matrices. While the CNN model demonstrated superior results, ViT and Xception models were evaluated for their potential applications, particularly as datasets grow larger.

The findings of this research have broader implications, as they highlight the potential for AI-based tools to support healthcare professionals and enhance early skin cancer diagnosis. Future directions include focusing on improving model performance for detecting skin cancer in diverse skin types, particularly those with darker skin tones, to ensure equitable healthcare access. The study concludes that while CNNs currently outperform transformer-based models with limited data, expanding datasets could unlock the full potential of ViTs and other emerging architectures.

1. Introduction

Did you know that skin is the biggest organ in the human body? It is soft, to allow movement, but still tough enough to resist breaking or tearing. Skin cancer represents a single risk for the skin.

Skin cancer rates are escalating globally at an alarming pace, marking a significant public health concern. The surge in these rates can be attributed to a multitude of interconnected factors. Firstly, the pervasive depletion of the ozone layer, a consequence of industrialization and environmental changes, has led to an influx of harmful ultraviolet (UV) radiation reaching the Earth's surface. This intensified exposure significantly heightens the risk of skin cancer, particularly melanoma, which is directly linked to UV radiation. Moreover, the changing dynamics of modern lifestyles contribute to increased sun exposure, with individuals engaging in outdoor activities without adequate protection. As populations continue to grow, urbanize, and adopt Westernized lifestyles, the incidence of skin cancer is expected to rise. Understanding and mitigating these contributing factors are crucial steps in developing effective strategies for early detection and prevention, emphasizing the urgency and importance of advancements in skin cancer detection technologies. Beyond the challenges posed by the escalating rates of skin cancer, a compelling driver for transformative change lies in the rapid advancements of deep learning and computer vision technologies. The integration of artificial intelligence (AI) has revolutionized medical diagnostics, offering unprecedented capabilities for skin cancer detection.

Imagine a future where artificial intelligence becomes our ally in the urgent fight against skin cancer, actively contributing to early detection and potentially saving countless lives. In this transformative era of healthcare innovation, our focus turns to dermatology, specifically the realm of skin cancer detection. As we navigate through the complexities of current diagnostic methods, we recognize the escalating global incidence of skin cancer cases, underscoring the critical need for advanced, efficient, and accessible detection techniques. Amidst this backdrop, the integration of cutting-edge technologies, particularly Convolutional Neural Networks (CNNs) and Vision Transformer models (VITs), emerges as a beacon of hope.

I will be working with two doctors, Dr. Melina Nikaki and Dr. Rafaela Argyriadi. Dr. Melina specialized in Dermatology - Venoresiology entirely at the Hospital for Dermatology and Venereal Diseases "A.Sygros" attending and actively participating in all the specialized departments and clinics of the Hospital, acquiring knowledge in the entire spectrum of Clinical Dermatology, Sexually Transmitted Diseases, Pediatric Dermatology, Interventional Dermatology. She gave me a long illustration about the procedure for skin cancer detection and detection for skin lesions in general.

Dr. Rafaela operates as a sub-investigator in the cancer department at the University Hospital of Patras. She advised me throughout the duration of the project and shared her knowledge about oncology.

This initiative is really important to me because I want to be able to help people, especially with such a delicate subject. I was speaking with Dr. Melina when she told me that during the pandemic of covid-19 some doctors specializing in dermatology started medical examination from home, In order for the doctor to examine the patients' skin, the patients would either send pictures of their skin lesions or hold online sessions. Therefore, my own effort in this project is to see if it would be possible to create models that can perform this function. If it is not possible for the patient to see a doctor at that time, especially in remote areas or on our Greek islands where doctors are few, they will be able to provide a first assessment.

In summary, the identification of skin cancer is an important problem with broad ramifications for world health. The goal of this research is to develop prediction models and evaluate which ones are most appropriate for the initial phases of the skin cancer detection procedure.

2. Terms of reference/objectives and literature review/desktop research

The goals of the project and a review of the literature are described in this chapter. The research findings center on the identification of skin cancer. The main focus of the study topics is how to improve or simplify the process for people who have skin cancer or who have suspicions about developing the disease.

2.1 Terms of Reference/Objectives

2.1.1 Research Statement

The requirement for people to have access to healthcare at all times, especially in the absence of close doctors, serves as the inspiration for the research statement.

2.1.2 Research Questions

The project's goals and objectives lead to the primary research question, which is: "How can I help people with this whole skin cancer detection process?" How can I improve their lives, particularly for those who reside in areas without access to medical care or hospitals? Both of the follow-up query explore the project's larger framework and ask, "How far can we go with the technology we have available?"

2.1.3 Objectives

The researcher's goals, which are mine, include creating various models by using data from HumanAgainstMachine (HAM10000).

My goal with these models is to classify the dataset's seven distinct forms of skin cancer. Additionally, assess whether it is possible to employ these models for the detection of skin cancer by comparing their accuracy.

The main goal of the research is to construct a classification model using advanced technology. We will additionally acquire supplementary data regarding demographic attributes from the Exploratory Data Analysis phase.

The dissemination of research-derived actionable insights to stakeholders is one of the main goals.

2.2 Literature Review/Desktop Research

2.2.1 Overview of Key Themes in Skin Cancer Detection

The literature surrounding AI in skin cancer detection is rich and varied, focusing on enhancing early diagnostic accuracy through deep learning techniques. Several studies have explored the use of Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and other machine learning methods to improve image-based diagnosis. Dildar et al. (2021) emphasized that CNNs demonstrated superior performance in melanoma detection, achieving up to 97.5% accuracy, which aligns closely with the outcomes of this project. Their findings underscored the potential of non-invasive and cost-effective deep learning approaches for rapid diagnostics.

2.2.2 Comparing AI Techniques and Their Clinical Utility

While CNNs remain a gold standard in image classification for medical diagnostics, recent research, such as that by Lyakhova & Lyakhov (2024), has introduced Vision Transformer models as promising alternatives. ViTs show potential in capturing global image features, offering a different approach from CNNs' localized feature extraction. However, ViTs require larger datasets to fully realize their potential, as also reflected in my findings. Similarly, Yang et al. (2024) evaluated the performance of both traditional machine learning techniques and advanced deep learning models, suggesting that improving model generalization and robustness is key for real-world application.

2.2.3 Addressing Challenges: Data Quality and Diversity

A common theme in the literature is the challenge of data diversity and imbalance. Furriel et al. (2024) highlighted the need for AI models that consider diverse demographics and address data standardization issues to ensure reliable diagnostics across populations. This is particularly relevant when applying AI in dermatology, where lesion characteristics may differ based on skin type, age, and gender. The findings in my research echo this sentiment, especially as model accuracy varies significantly with the size and variety of training data.

2.2.4 Implications for Future Research and Practice

The literature consistently points to the importance of developing AI models that are clinically validated and can be translated effectively into healthcare settings. With increasing interest in mobile health applications and remote diagnostics, the studies reviewed suggest a growing trend towards integrating AI-based skin lesion detection into everyday healthcare. The key takeaway is that while the technology holds promise, achieving optimal performance requires addressing data quality issues, improving model interpretability, and working closely with healthcare professionals to ensure these tools are practical and user-friendly.

2.2.5 Action Research in Healthcare and AI Development

Overview of Action Research

Action research is a participatory and reflective methodology where the researcher engages in problem-solving while simultaneously studying and improving the process. In healthcare, action research is particularly relevant because it allows for iterative improvements based on practical, real-world feedback. Its cyclical process of planning, acting, observing, and reflecting is well-suited for contexts requiring continuous development and adaptation, such as in healthcare and AI applications.

Relevance to Healthcare and AI Development

In healthcare contexts, action research has been widely used to enhance clinical practices, patient outcomes, and healthcare operations. For instance, Bradbury et al. (2019) highlight how action research bridges practice-based evidence with practical solutions, enabling changes to be tested and refined in real-world settings. When applied to AI model development, this methodology supports continuous model improvement through iterative cycles of experimentation, evaluation, and revision. For example, training and validating AI models, as presented in this thesis, mirrors the "action" and "reflection" stages, and working closely with clinicians emphasizes the "participatory" nature of action research.

Connections to AI and Skin Cancer Research

Action research is particularly suited for AI model development in healthcare as it allows for iterative learning and adaptation based on real-world feedback from medical practitioners. In the context of skin cancer research, this approach enables the refinement of AI models based on the insights provided by clinicians throughout each development cycle. Each phase of model evaluation and improvement becomes an opportunity to reflect, adapt, and enhance model accuracy and applicability, aligning with action research principles.

Applying Action Research in This Project

In this thesis, the action research methodology was applied throughout the development and evaluation of AI models for skin cancer detection. The iterative cycles involved planning and building multiple models (e.g., CNN, ViT, Xception), followed by rigorous testing and validation. Each phase involved close collaboration with medical professionals, who provided insights and feedback that were used to refine the models' performance and clinical relevance. This participatory approach ensured that the models were not just technically sound but also aligned with practical healthcare needs. The reflective process, where outcomes were reviewed and adjustments were made, further allowed the research to adapt to challenges such as data limitations and model accuracy issues.

This approach exemplifies the core of action research by engaging in continuous cycles of action, reflection, and adaptation, thereby enabling the successful development of AI models tailored to the complexities of skin cancer diagnosis.

Explicitly Relating Action Research Cycles to Project Cycles

Throughout this project, the cyclical nature of action research was mirrored in the phases of model development. The planning phase involved strategizing which deep learning models would be tested and how they would be validated. The action phase included training and refining models like CNN, ViT, and Xception. The observation phase was characterized by careful analysis of model performance metrics, and the reflection phase incorporated feedback from medical professionals to improve model generalizability and accuracy. This continuous cycle of planning, acting, observing, and reflecting ensured that the AI models evolved in alignment with both clinical needs and technical standards.

3. Methodology

The project will investigate the early identification of skin cancer. The development of models capable of highly accurate image classification is my main goal. Let me first explain the approach and procedure for detecting skin cancer or lesions in general.

In the beginning, the doctor obtains a medical history from the patient, asking questions such when the mark was first noticed or when the patient last remembered it. Then the doctor examines the mark or may want to use a magnifying glass. Next, if the doctor believes that the mark is suspicious, he uses a traditional dermatoscope or moves on to digital dermatoscopy. This equipment provides the doctor with a better and clearer image of the patient's mark by magnifying far more than a standard magnifying glass. If the dermatoscopy process reveals any indications that the mark is worrisome, the doctor will attempt to save a small portion of the skin and send it for a biopsy. The skin mark will be removed if the biopsy confirms that the mark is cancer.

The methodology of this project was deeply rooted in action research (AR) principles, which focus on iterative cycles of planning, acting, observing, and reflecting. As Bradbury and Reason (2019) describe, AR enables the researcher to actively engage in problem-solving while simultaneously studying the process. This cyclical nature of AR was particularly suited for developing AI models for skin cancer detection, as it allowed for continuous reflection and adaptation based on real-world feedback from medical experts. Each research cycle began with planning—defining objectives for model development—followed by action through model training and refinement. Observations from performance metrics were then reflected upon with medical professionals, guiding subsequent adjustments. This continuous, reflective process ensured that the models evolved not only to meet technical standards but also to align with clinical needs, thus adhering to the core principles of action research.

After viewing the medical procedure, let's briefly discuss the technologies we will be using before moving on to the dataset.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a class of deep learning algorithms that have revolutionized the field of computer vision. Introduced by Yann LeCun in the late 1980s and early 1990s, CNNs gained widespread popularity due to their success in image recognition tasks, such as the ImageNet competition. The architecture of CNNs is designed to automatically and adaptively learn spatial hierarchies of features through backpropagation by using multiple building blocks, such as convolution layers, pooling layers, and fully connected layers.

CNNs are particularly known for their ability to capture spatial and temporal dependencies in an image through the application of relevant filters. This ability makes them highly effective for tasks like image classification, object detection, and image segmentation. The key innovation in CNNs is the convolutional layer, which uses a set of learnable filters that apply convolution operations across the input image to detect features such as edges, textures, and shapes at different spatial hierarchies. Over time, CNN architectures have evolved with models like AlexNet, VGGNet, ResNet, and Inception, each contributing improvements in depth, training efficiency, and accuracy.

Vision Transformers (ViTs)

Vision Transformers (ViTs) represent a newer approach in the field of computer vision, leveraging the transformer architecture originally developed for natural language processing tasks. Introduced by researchers at Google in 2020, ViTs have shown that transformers, which rely on self-attention mechanisms, can outperform traditional CNNs when scaled appropriately and trained on large datasets.

The core idea behind Vision Transformers is to treat an image as a sequence of patches, similar to how words are treated in a sentence for NLP tasks. Each image is divided into fixed-size patches, which are then linearly embedded, combined with positional encodings, and fed into a standard transformer encoder. The self-attention mechanism within transformers allows ViTs to capture long-range dependencies and contextual relationships between different parts of the image, leading to improved performance on various vision tasks.

ViTs have gained recognition for their flexibility and robustness, particularly in tasks that require understanding complex patterns and relationships within an image. Studies have demonstrated the superiority of ViTs in several benchmarks, achieving state-of-the-art results in image classification, segmentation, and object detection

Now that we know the doctors' statements and the technologies, we know the dataset needs a large number of high-quality images, which is why we selected the Human Against Machine (HAM10000). A huge collection of 10.000 images showing seven distinct cancer types:

- akiec : Actinic Keratoses and intraepithelial carcinoma / Bowens Disease.
- Bcc : Basal Cell Carcinoma
- Bkl : Benigh Keratosis – line lesions
- Df : Dermatofibroma
- Mel : Melanoma
- Nv : Melanocytic Nevi
- Vasc : Vascular lesions

The research methodology will be conducted as follows:

Data Cleaning and Analysis: After cleaning the dataset, we will use exploratory data analysis to extract a variety of insights from the actual data.

Model Development: we will create 3 different models

- CNN Convolutional Neural Network
- VIT Vision in Transformers
- Xception

we will perform cross-validation to obtain the best accuracy possible.

Model Comparison: I will evaluate the accuracy of the models and see whether this project can aid in the early detection of skin cancer.

However, there are certain limitations to my methodology. Which might contain more technologies. However, the dataset is another restriction, only the HAM10000 was determined to have very high-quality data. My understanding of dermatology may also be a limitation, however I aim to improve over this project.

4. Project Activities in Cycles

I will go over the cycles and actions done to make this project a reality in this chapter. I'll go over the planning and analytical techniques I used to make sure I obtained the desired outcomes. In addition, I'll discuss how each stage put us on distinct routes and provide a final review.what I accomplished, providing a comprehensive view of the project's development. The cyclical nature of the project's development mirrored the core tenets of action research. During **Cycle 1**, the foundation of the project was laid through discussions with experts and preliminary data analysis, which aligns with AR's emphasis on planning and collaboration. **Cycle 2** saw the iterative development and refinement of models like CNN, ViT, and Xception, embodying the AR principle of acting and observing in cycles. At each stage, the reflective input from Dr. Melina Nikaki and Dr. Rafaela Argyriadi was crucial, as Action Research stresses the importance of engaging stakeholders to ensure the research's relevance and applicability. This allowed the project to remain adaptable and responsive, a key feature of the AR approach in healthcare settings (Bradbury et al., 2019).

4.1 Cycle 1: Foundations and Initial Analysis

- **Constructing:**
  The construction phase involved establishing the theoretical framework and selecting a compelling healthcare issue—skin cancer detection. Identifying and understanding the scope of the problem, along with initial discussions with experts like Dr. Melina Nikaki, laid the groundwork for setting research objectives.

- **Planning:**
  Planning focused on formulating a research proposal and defining the data collection process. This included sourcing a high-quality dataset (HAM10000) and outlining the technical goals for developing predictive models. Collaboration with Dr. Rafaela Argyriadi was crucial during this phase to ensure medical relevance and practical feasibility.

- **Taking Action:**
  Action in this cycle consisted of exploratory data analysis (EDA) to comprehend the dataset's distribution and characteristics. Cleaning and visualizing data patterns facilitated a better understanding of lesion types, demographics, and diagnosis trends. Using Python and relevant libraries, initial insights were extracted to shape the development of AI models.

- **Evaluating Action:**
  The evaluation of EDA outcomes was conducted in close consultation with medical experts, validating the accuracy of the insights gained. This phase informed the development of AI models and provided a critical checkpoint for ensuring that technical progress aligned with healthcare objectives.

4.2 Cycle 2: Development and Refinement

- **Building the Product:**
  During this cycle, model development began in earnest. A CNN was first constructed due to its efficiency with image data. ViT models were subsequently explored to test their global feature extraction capabilities. Finally, the Xception model was built to compare against CNNs and transformers, with the aim of developing a robust solution for skin cancer classification.

- **Updating the Product:**
  The iterative refinement of models was a key feature of this cycle. Each model underwent multiple rounds of training, fine-tuning, and validation. Techniques such as K-Fold cross-validation for CNN and layer freezing for ViT were used to optimize model performance. Continuous feedback from domain experts ensured that adjustments made to model parameters improved clinical relevance and technical efficacy.

- **Finalizing the Product:**
  The final stage involved rigorous model testing to validate accuracy and generalizability across unseen data. The CNN model, achieving the highest accuracy, was refined to demonstrate the project's success in developing an effective image classification tool for skin cancer. Additionally, discussions with medical experts were held to evaluate the potential of transforming this research into a practical diagnostic application for remote and underserved communities.

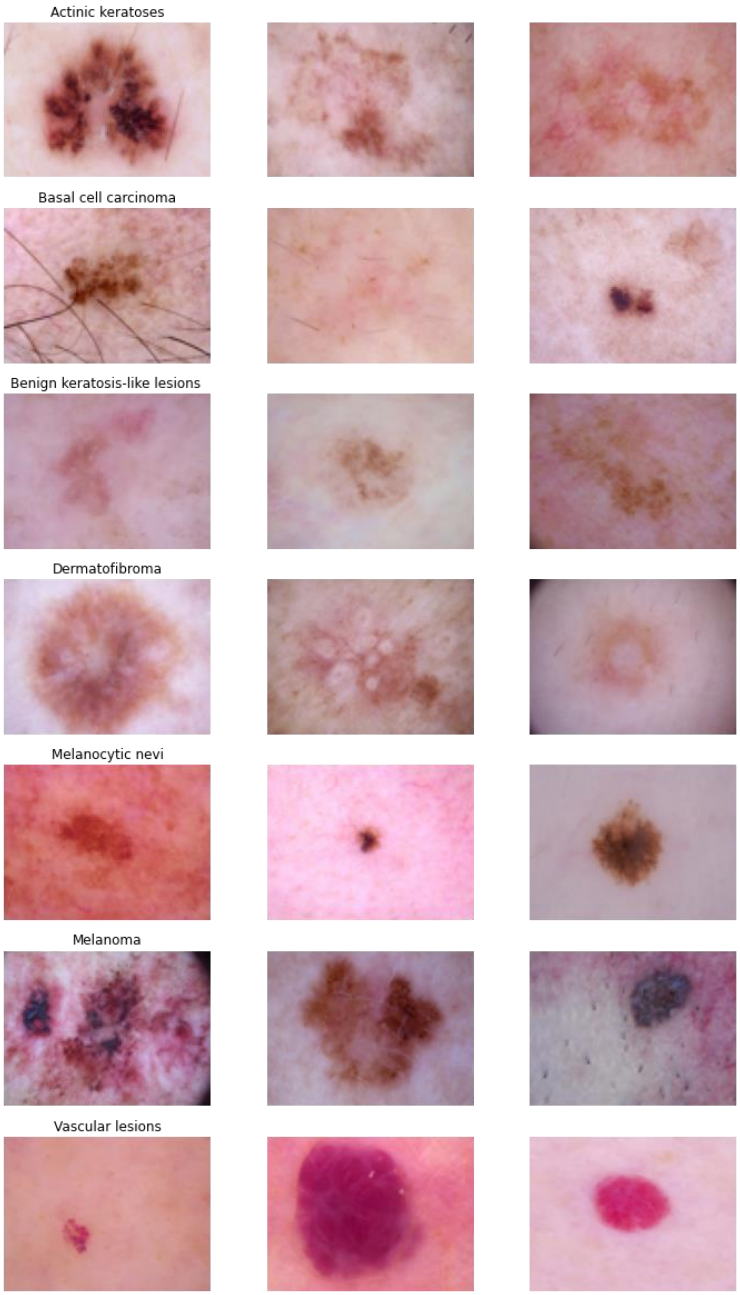
The action research cycles provided a reflective framework for continuous development and refinement of AI models for skin cancer detection. Each cycle—construction, planning, action, and evaluation—enabled iterative learning and model adjustments in response to feedback and observed results. This ongoing reflective process laid the foundation for the technical findings presented in the following section. Section 5 will discuss in detail the outcomes of these cycles, focusing on model performance, evaluation metrics, and insights gained from implementing and testing various deep learning architectures.

5. Project Findings

Initially, my goal was to extract as much information as possible from the dataset, including gender, age, and lesson collection distributions.

5.1 Exploratory Data Analysis

Here are the seven different types of cancer:



Actinic Keratoses and intraepithelial carcinoma / Bowens Disease

Basal Cell Carcinoma

Benign Keratoses-Line Lesions

Dermatofibroma

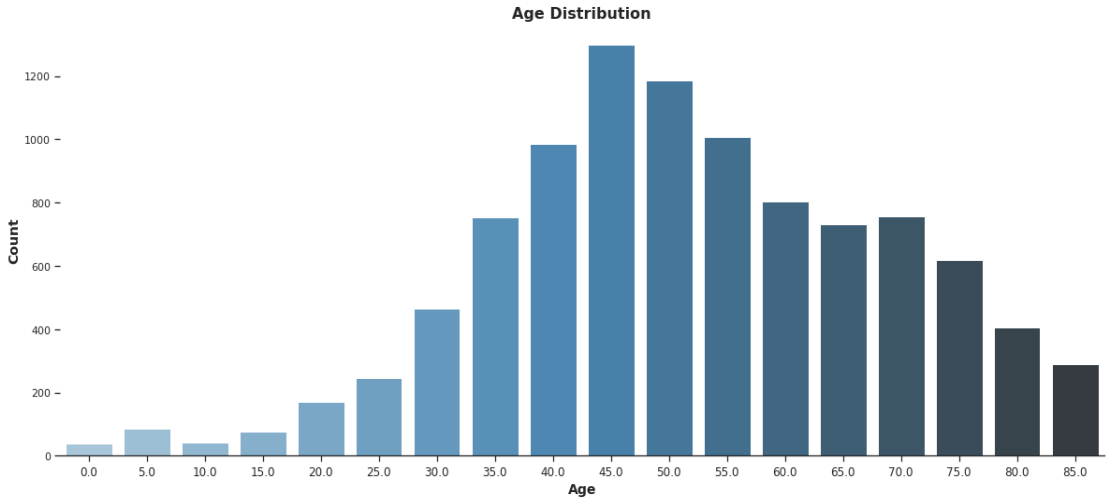
Melanoma

Melanocytic Nevi

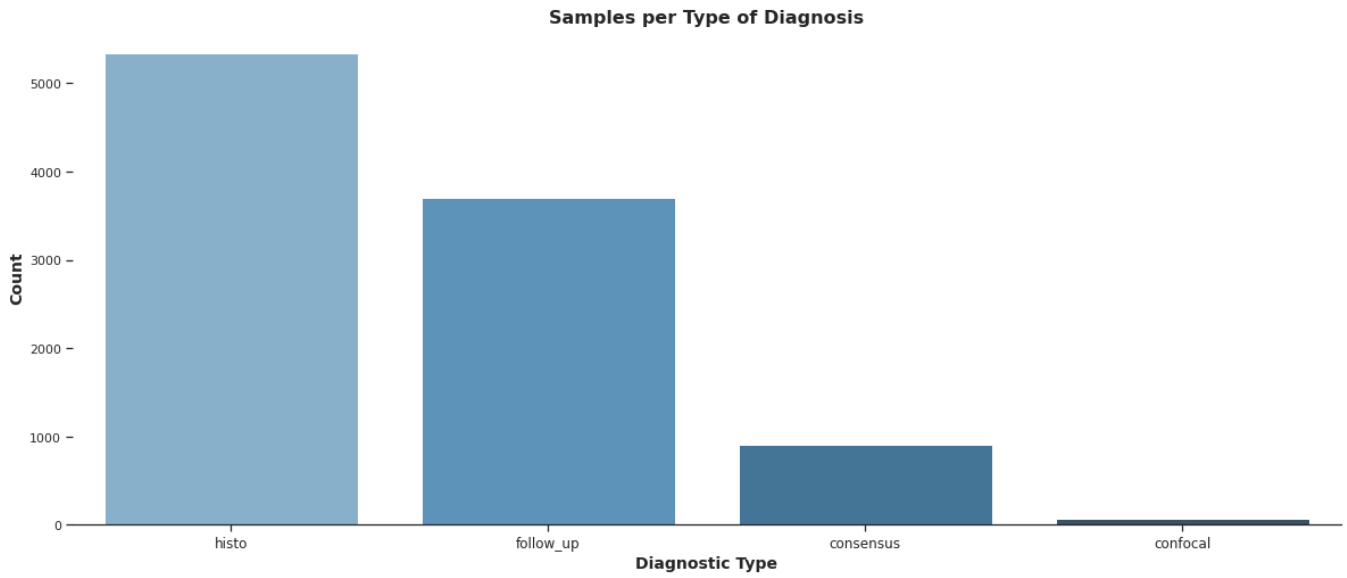
Vascular Lesions

name	dtypes	missing	uniques
lession_id	object	0	7470
image_id	object	0	10015
dx	object	0	7
dx_type	object	0	4
age	float64	57	18
sex	object	0	3
localization	object	0	15
path	object	0	10015
cell_type	object	0	7
cell_type_idx	int8	0	7

The summary helps to understand the type of metadata collected. We noticed that for some lesions there must be more than one image as the lesion and image ID do not match. The Unique columns also indicate the number of classes (dx = 7), and how the age, sex and localization features were organized.



The majority of the samples, as we can see, are from patients between the ages of 40 and 55. After the age of 25, the number of samples increases significantly, nearly doubling for those aged 30 and nearly doubling again for those aged 35. The number of samples stays virtually constant between the ages of 60 and 70, then starts to decline beyond 75.

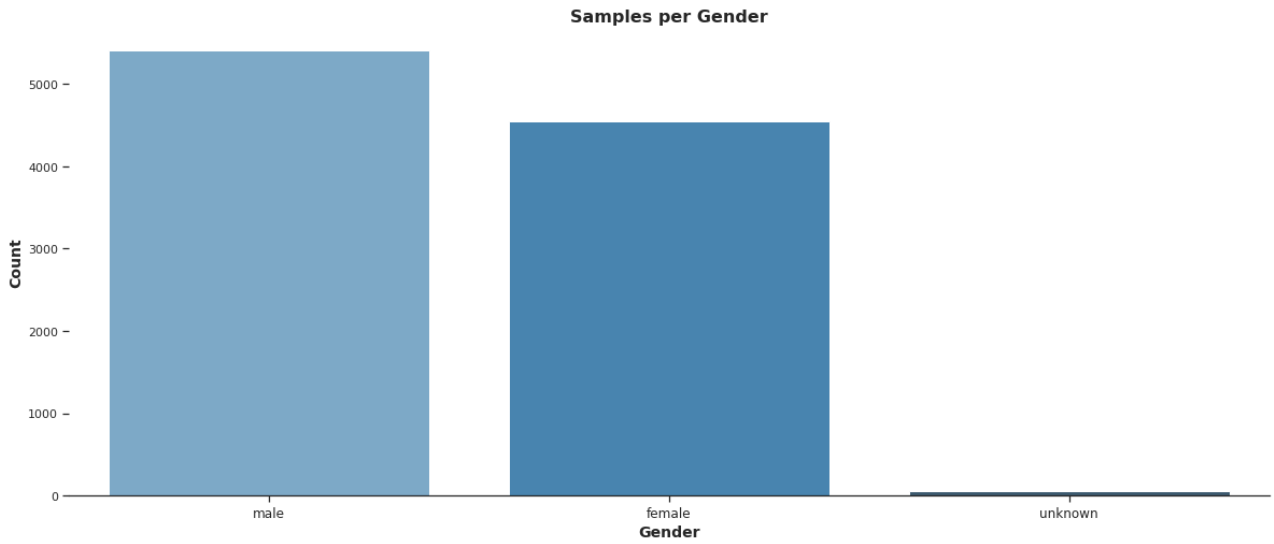


- Histo: The statistics indicates that, in 53% of the patient cases, skin cancer was discovered during the Histopathology. It’s the process that Dr Melina told us.

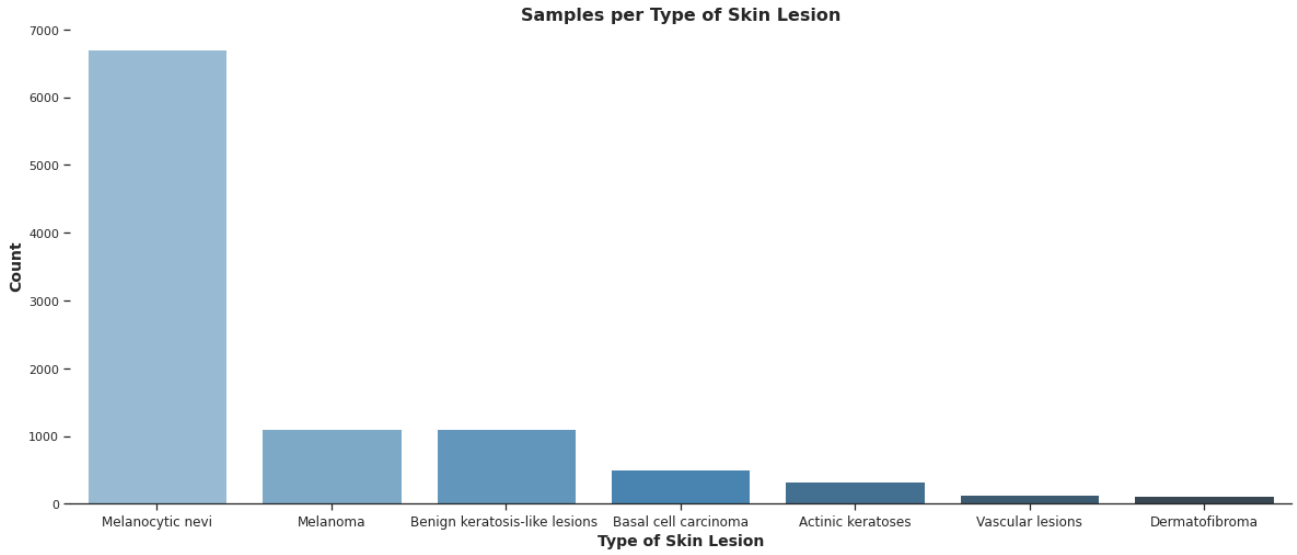
- Follow up: It has become the case that 37% of patients learned they had skin cancer through a follow-up exam. This procedure aims to track the patient's development, evaluate the efficacy of the care, and spot any new or persistent issues.

- Consensus: The 9% of the patients diagnosed with cancer with consensus, it is a process reached among experts in a particular field, this agreement usually results from discussions, analyses, and evaluations of available data and knowledge. The goal of expert consensus is to provide valid and reliable recommendations based on the experience and expertise of the participants.

- Confocal: The remaining 1% were diagnosed with cancer by the in-vivo Confocal procedure, an imaging procedure that allows examination of living tissue at the cellular level without the need for biopsy.



The participants are mostly males 54% and females 45% the rest 1% is unknown. There is no significant difference between the genders.



From the statistics we can see that:

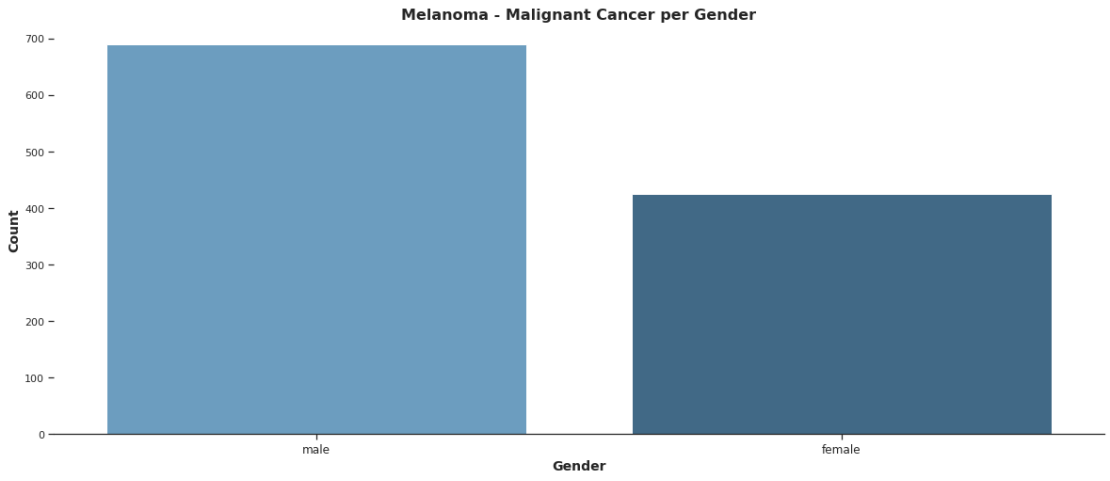
- Melanocytic Nevi: has the highest presence in the dataset about 67%.
- Melanoma: 11%
- Benign keratoses Lesions: 11%
- Basal Cell Carcinoma: 5%
- Bowens' Disease: 3%
- Vascular Lesions: 1%
- Dermatofibroma: 1%

In the dataset, Vascular lesions and dermatofibroma are present in 1% of cases each. Although we were aware of this when selecting the dataset, we still favored quality over quantity. Given that there are less images available for training the models for vascular lesions and dermatofibroma than for other types of lesions, we may anticipate that these lesions will not attain as high accuracy levels, although this is not granted confirmation.

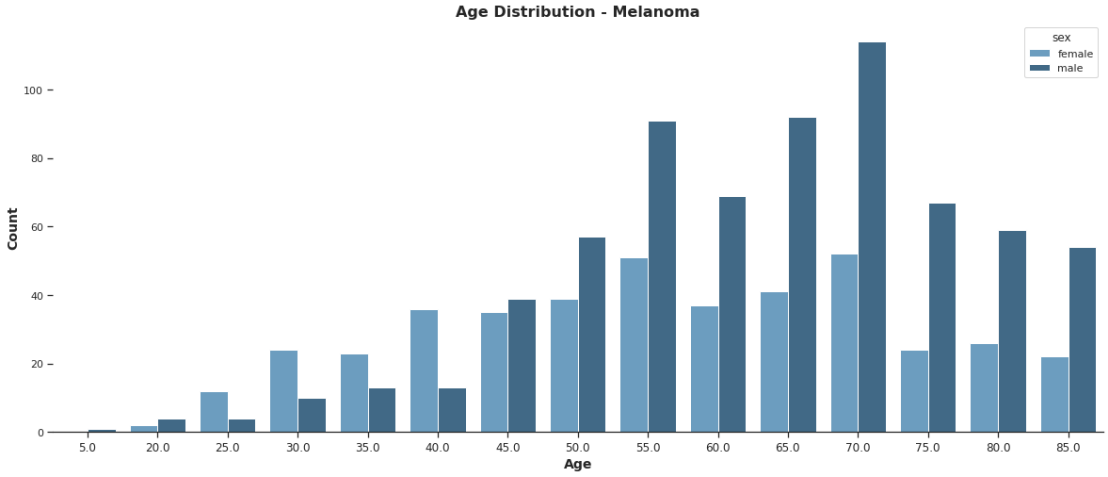
Presence of localization in the dataset:

- Back: 22%
- Lower extremity: 20%
- Trunk: 14%
- Upper extremity: 11%
- Abdomen: 10%
- Face: 7%
- Chest: 4%
- Foot: 3%
- Unknown: 2%
- Neck: 1%
- Scalp: 1%
- Hand: less than 1%
- Ear: less than 1%
- Genital: less than 1%
- Acral: less than 1%

The only malignant type of skin lesion present on this dataset is Malignant Melanoma, where surgical removal in the early stage of cancer can provide a cure. The remaining skin lesions are benign even though they may require treatment. That is the reason I carried out additional research on melanoma.

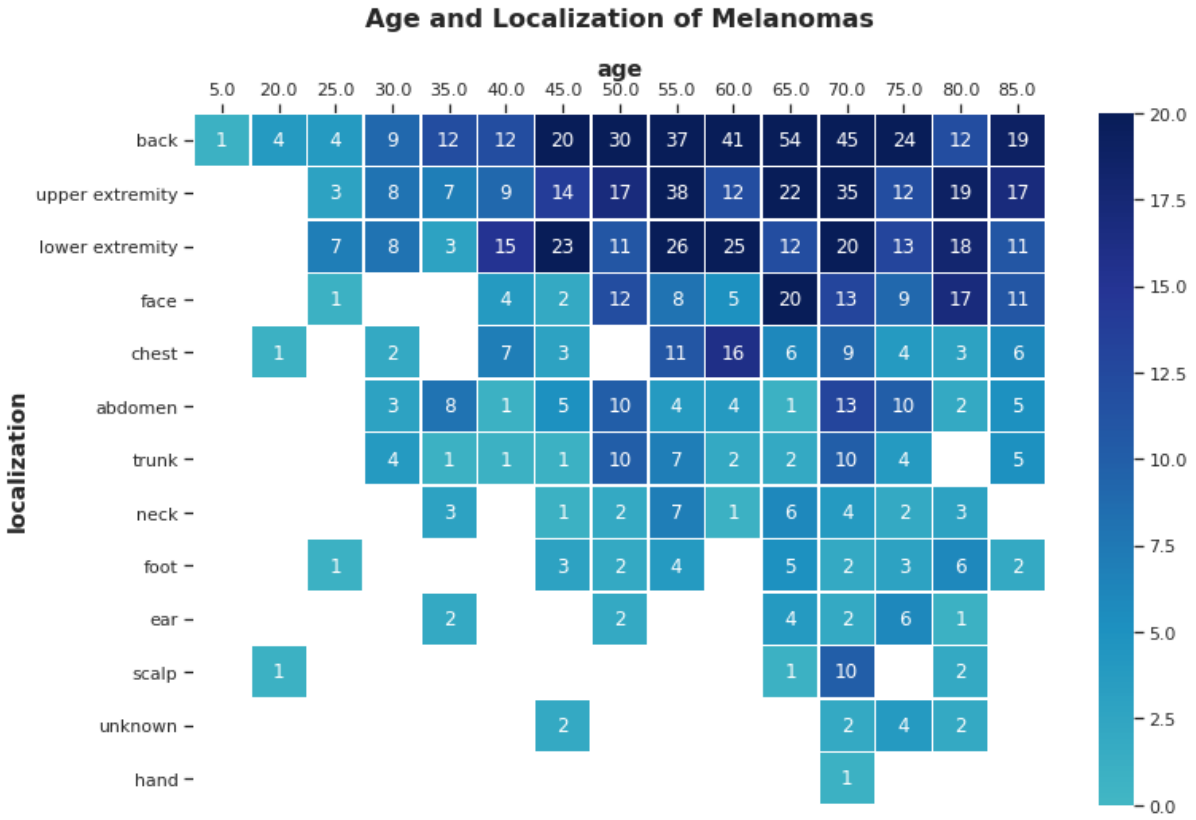


We observed a similar distribution across the two primary categories in terms of gender. But if we limit our analysis to melanoma, the above plot indicates that males account for more than 60% of reported instances.



The age distribution is quite different for Melanoma diagnostics when compared to the whole dataset.

There are two separate peaks, at 55 and 70 years old, for both genders. The peaks might correspond with when people often have comprehensive medical examinations. Females are more likely to have younger melanoma samples. Significant gender disparity between 25 and 40 year olds, Melanoma is more common in older people who are males. The number of instances involving female samples appears to stabilize at the age of 40. Compared to the male samples, the peaks at 55 and 70 years old are less expressive.



The heatmap makes a good representation of how age influences Cancer incidence. Note the cluster between the ages of 45 to 70. This cluster that we see is logical since most of the data are found in these ages and in these parts of the body. For the age group of 50 and 70 years old, the face, abdomen, chest and trunk also present a higher number of incidence. The scalp seems to be a more common localization only for 70 years old. The localizations do not seem to be related to the parts of the body most commonly exposed to the sun. If it was the case, scalp, hands and face should have a higher incidence.

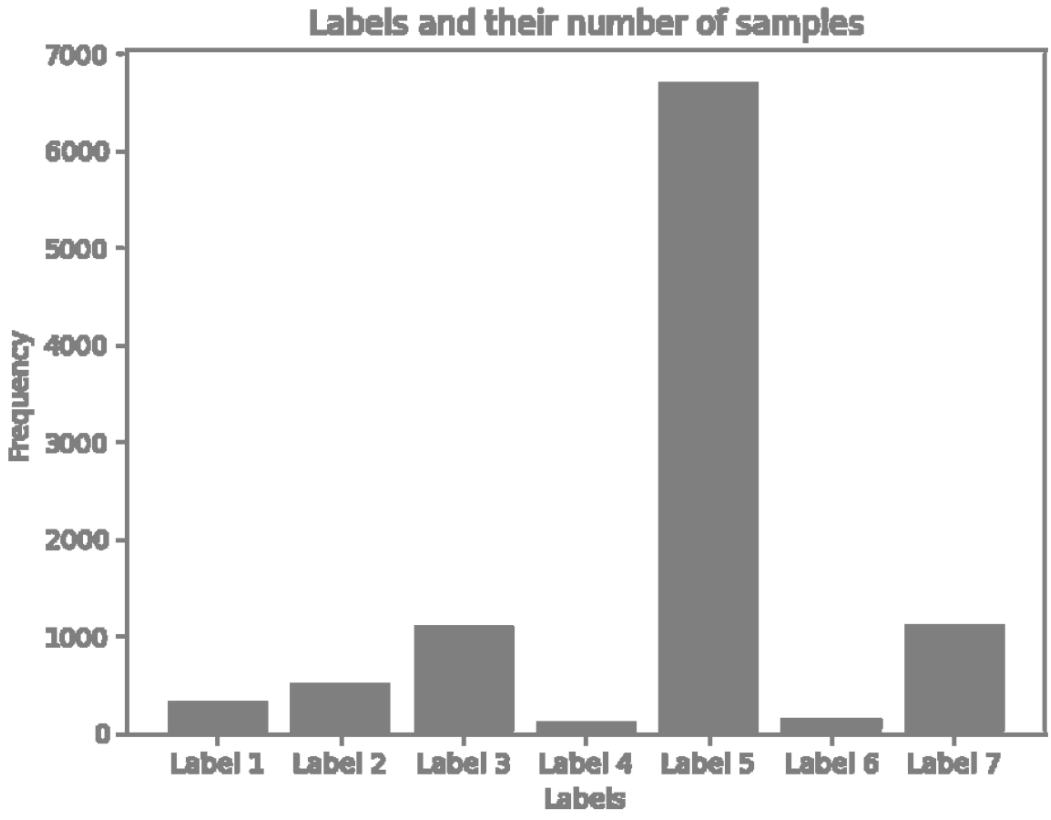
5.2 Classification Models

I performed some classification investigation to develop models that can classify images based on the above findings.

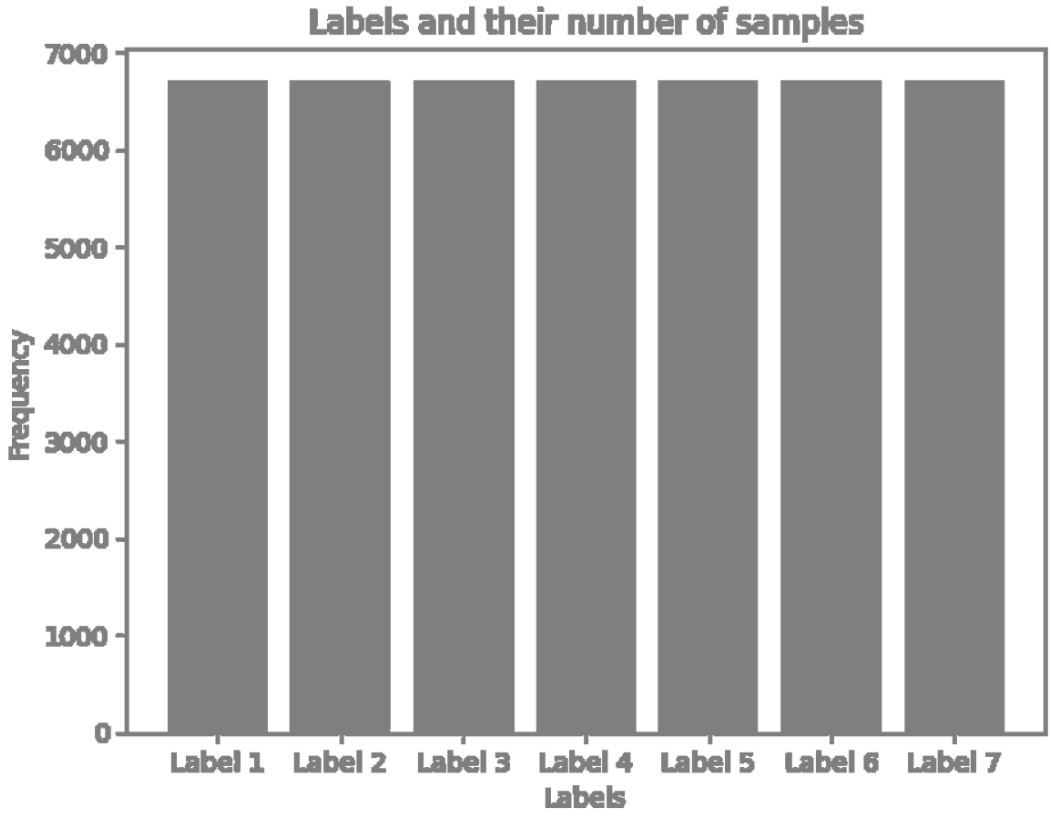
5.2.1 Convolutional Neural Network

First of all I had to do some oversampling, Skin cancer datasets are often imbalanced, meaning certain types of skin lesions are underrepresented. Oversampling helps balance the dataset, ensuring the model doesn't become biased toward more frequent labels. This is crucial for medical diagnostics to avoid false negatives for less common but dangerous conditions like melanoma.

Dataset before resampling:



Dataset after resampling:



After that, I had to reshape the images and normalize them. The images are reshaped to a 4D shape (28, 28, 3) to match the input expected by the convolutional layers. Normalizing pixel values (dividing by 256) is important for stabilizing training and ensuring the model doesn't encounter extreme values, making it easier for the network to learn.

CNN Model Architecture

Convolutional Layers:

We used multiple convolutional layers with filter sizes of 32, 64, 128, and 256 filters. The reason for this structure is that convolutional layers are designed to extract spatial features from the images, such as edges, colors, and textures. The use of multiple convolution layers allows the model to progressively capture more complex features, such as lesion borders and patterns, which are critical for accurate skin cancer detection. By gradually increasing the number of filters, we ensured that the model could identify both simple and intricate patterns from the images, which are necessary for distinguishing between different types of skin lesions.

Batch Normalization:

To improve the model's training stability and convergence speed, batch normalization was used after several layers. This technique normalizes the outputs of each layer, which makes the network less sensitive to the initial weights and learning rate. This also helps prevent overfitting, allowing the model to generalize better when applied to unseen data.

MaxPooling Layers:

MaxPooling layers were used to reduce the spatial dimensions of the image, which decreases the computational cost and helps to prevent overfitting. By retaining only the most important features, MaxPooling ensures that the model focuses on the most critical aspects of the image while reducing the risk of overfitting due to too much spatial detail.

Dense and Dropout Layers:

After feature extraction through convolutional layers, the data is flattened and passed through dense layers to classify the image into one of the seven categories. Dropout layers are included to prevent overfitting by randomly disabling some neurons during training, ensuring the model doesn't memorize the training data and generalizes well to new, unseen data.

Output Classes:

The model is designed to classify skin lesions into 7 categories, which aligns with medical classification standards. This decision was made in collaboration with doctors to ensure that the model outputs would be meaningful and practical for clinical use. The categories represent various types of skin lesions that are relevant to diagnosis in dermatology, making the model more clinically applicable.

Model Performance Across K-Folds

For training and evaluation, the CNN model was subjected to K-Fold cross-validation with 5 splits. This method ensured a robust assessment of the model's performance, preventing overfitting to a specific data partition and providing a more accurate measure of how the model performs on unseen data.

Training vs. Validation Accuracy:

Across the different folds, the training accuracy consistently reached near 100% within a few epochs, indicating that the model is learning well from the training data. The validation accuracy, while slightly lower than training accuracy, still demonstrated strong performance across all folds. The validation accuracy curves show slight fluctuations, which is typical when training on different subsets of data.

Training vs. Validation Loss:

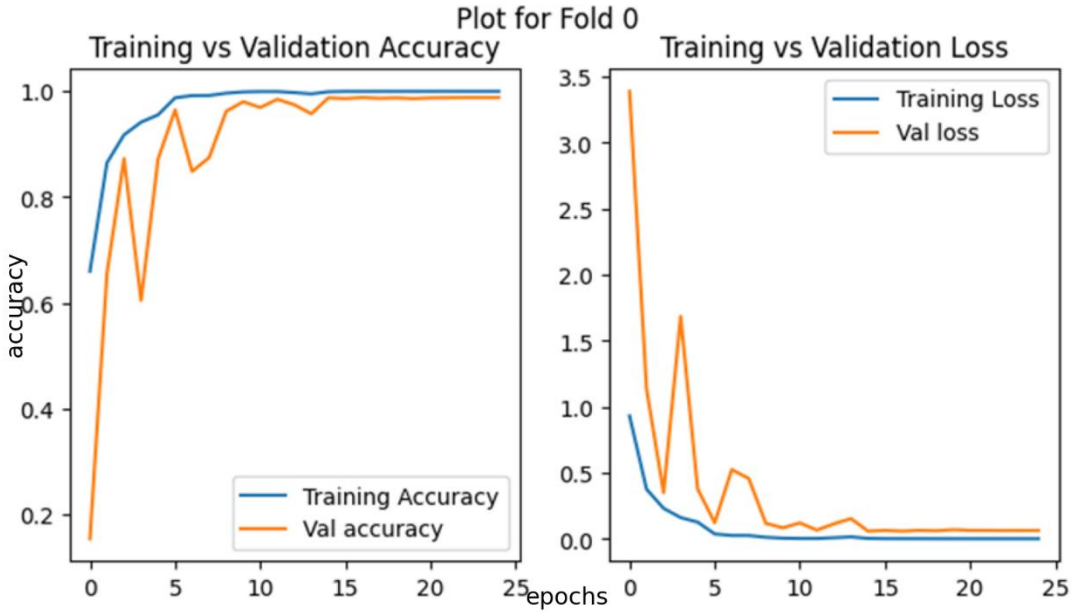
The training loss decreased rapidly in the early epochs, showing that the model is effectively minimizing error on the training data. The validation loss, though slightly higher than the training loss, converged over time, reflecting the model's ability to generalize to new data. In some cases, the validation loss fluctuated slightly, suggesting potential overfitting on certain folds, but the model still showed good generalization overall.

K-Fold Cross-Validation:

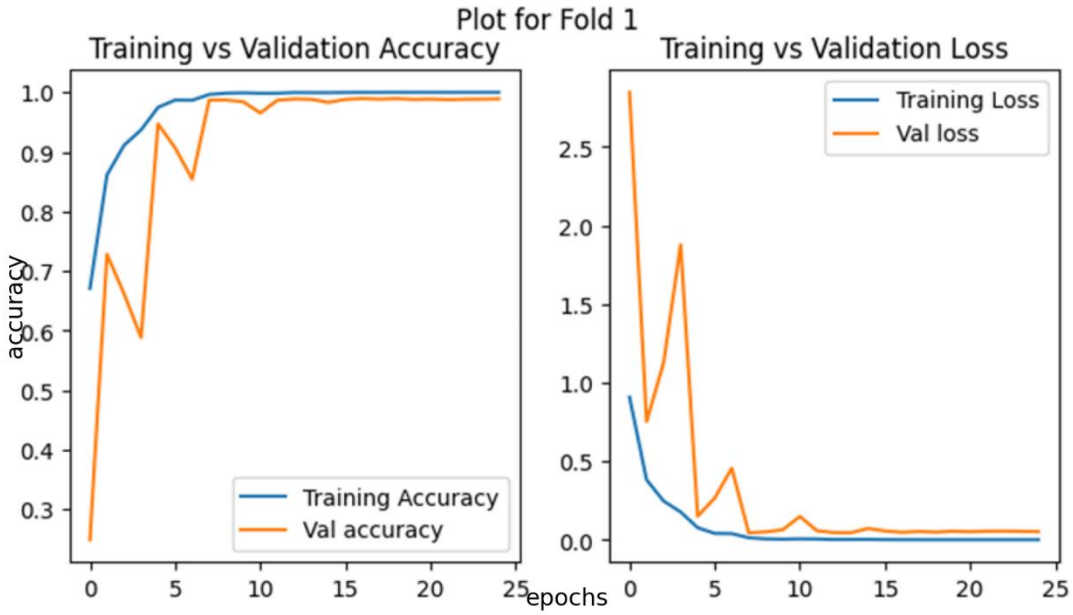
The use of K-Fold cross-validation allowed the model to be trained and tested on different subsets of data, leading to a more reliable evaluation of its generalizability. The performance across all folds remained consistent, reinforcing the robustness of the model architecture.

1. **Training vs Validation Accuracy:**
    - The blue line represents the accuracy achieved by the model on the training data across 25 epochs.
    - The orange line shows the accuracy on the validation set during each fold.
    - These graphs help visualize how the model's ability to classify skin cancer images improves over time. Ideally, both the training and validation accuracy should increase and converge, indicating that the model is learning well and generalizing to unseen data.
2. **Training vs Validation Loss:**
    - The blue line represents the loss (error) on the training set across the epochs.
    - The orange line shows the loss on the validation set.
    - The loss graphs depict how the model's prediction error decreases over time. A significant drop in the first few epochs indicates that the model is learning quickly. However, the gap between training and validation loss is important for spotting overfitting.

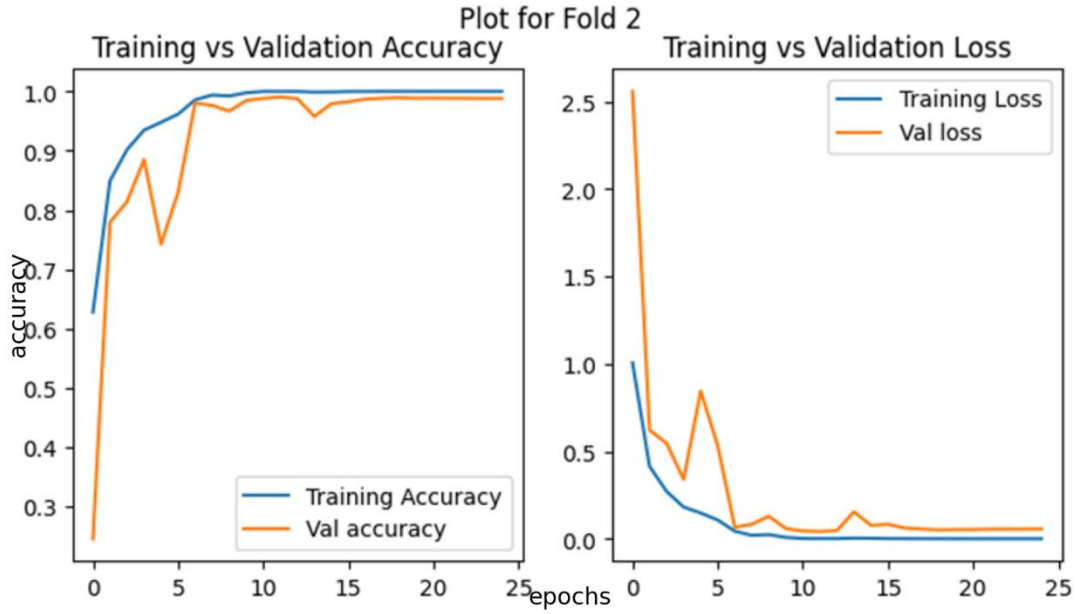
These graphs across the five folds give a clear picture of how the model performs on different subsets of the data, ensuring it generalizes well. You can see if the training and validation curves are close, which indicates that the model is not overfitting and performs consistently across multiple data partitions.



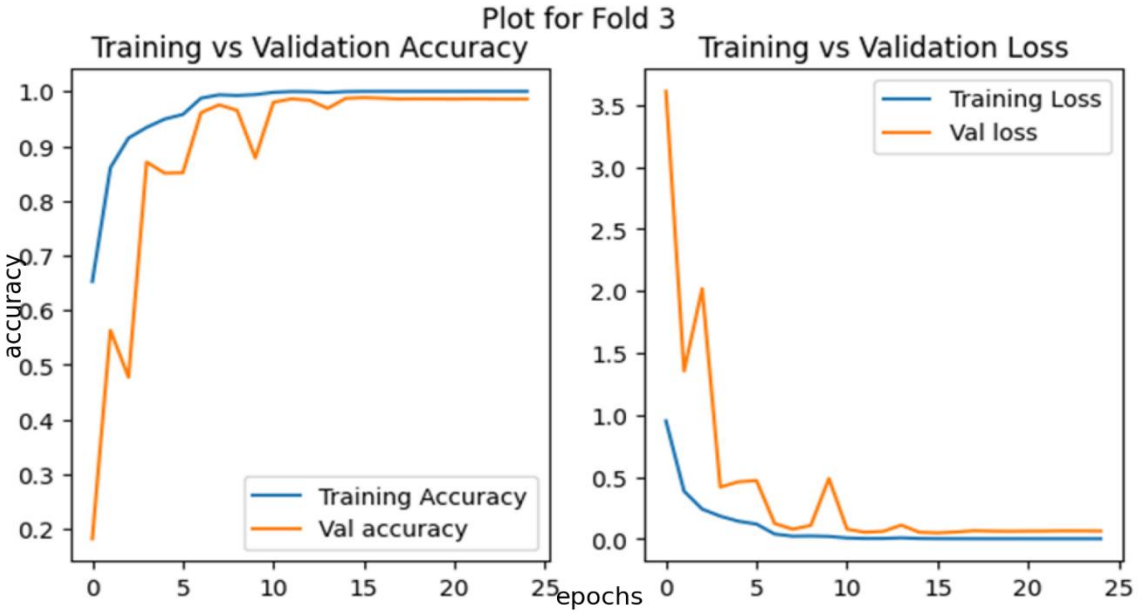
Test Accuracy: 98.82%



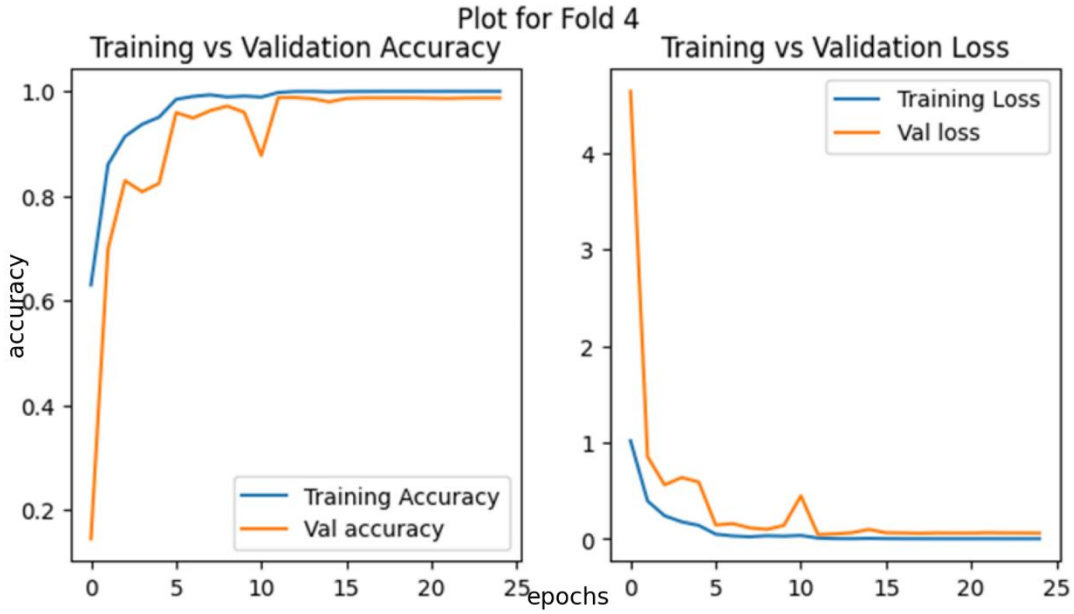
Test Accuracy: 98.90%



Test Accuracy: 98.80%



Test Accuracy: 98.59%



Test Accuracy: 98.79%

In this model, we used **accuracy** as a performance metric during the K-fold cross-validation process, despite the fact that accuracy is not the most reliable metric, especially in cases where data is imbalanced. However, accuracy was chosen because it provided a clear and simple understanding of the model's performance for the doctors we collaborated with. Since they were more familiar with accuracy as a metric, this decision facilitated better communication and interpretation of the results from a clinical perspective. It helped bridge the gap between data science concepts and medical insights, ensuring that the doctors could easily follow the model's effectiveness.

While accuracy was used as the primary metric for evaluating model performance, it is not always the most appropriate measure, particularly when dealing with imbalanced datasets like those encountered in skin lesion classification. Metrics such as precision, recall, and F1-score provide a more nuanced understanding of model performance by focusing on false positives and false negatives. These metrics are crucial in medical diagnostics, where the cost of misclassifications can be high. Thus, although accuracy was easier for clinical collaborators to interpret, future evaluations should also consider other metrics to provide a comprehensive performance assessment.

- High Accuracy: The model seems to learn the features of the data quite well, as evidenced by its high accuracy across all folds.
- Stability: The model is stable and dependable because there aren't many performance variations across folds.
- Low Learning Rate: The model is getting close to optimum performance, and there is little room for future improvement, as evidenced by the learning rate dropping to extremely low values (1e-05).

  For image classification, the CNN model works incredibly well, exhibiting great accuracy and stability throughout all k-fold cross-validation folds. This implies that the model can sustain low loss and generalize well to new data.

5.2.2 Vision in Transformer

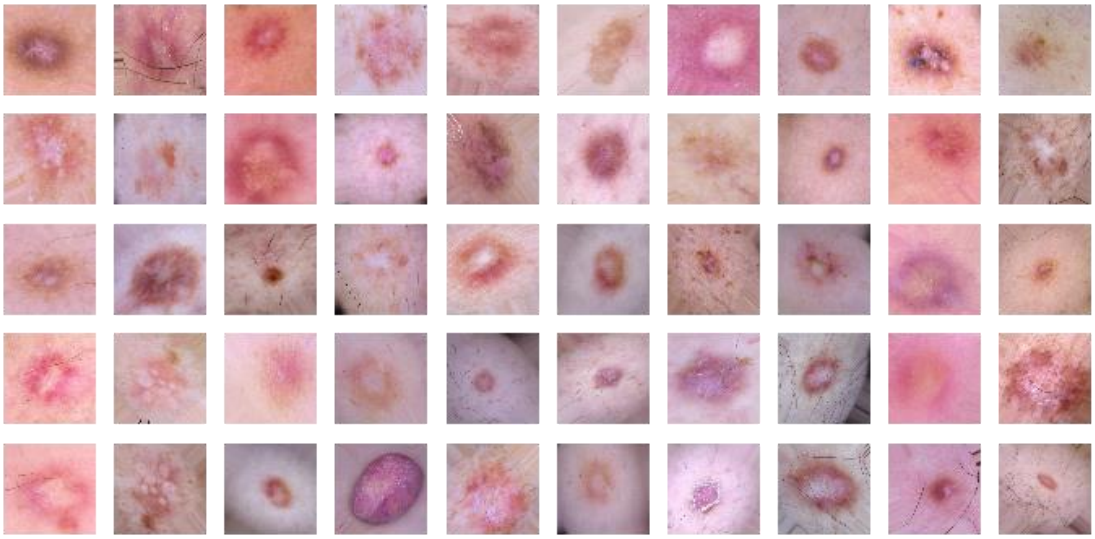
5.2.2.1 Vision in Transformer Model B 32

Image Resizing and Preprocessing:

The ViT B32 model was pretrained with images of size 224x224, and we resized the dataset images to this dimension for consistency with the model's input requirements.

ImageDataGenerator was used for real-time data augmentation, applying random transformations like rotations, flips, and zooms to prevent overfitting.

Augmented Images



Data augmentation is a critical pillar in the development of powerful and effective machine learning models, especially in the field of image processing. By increasing the amount of training data and improving the generalization of the models, it helps prevent overtraining and prepare the models to deal with a variety of real-world conditions.

Data Splitting:

We split the dataset into training (90%) and validation (10%) sets, ensuring that no duplicated lesions appear in both sets.

A class imbalance was observed, so we applied data augmentation on classes with fewer samples to reach a minimum of 6000 images per class.

Model Layers:

Vision Transformer (ViT) B32 Pretrained Model:

The core of our architecture is the ViT B32, which uses image patches and self-attention mechanisms to process the images. It was pretrained on ImageNet, and the top layer was excluded for custom classification on the 7 skin lesion classes.

This model captures long-range dependencies and is particularly suited for image classification tasks with its self-attention layers.

Flatten Layer:

The ViT output, which is a sequence of features, is flattened to a 1D vector before passing it to the fully connected layers.

Batch Normalization Layers:

Batch normalization layers are added after the Flatten layer and the Dense layers to improve training speed and stability.

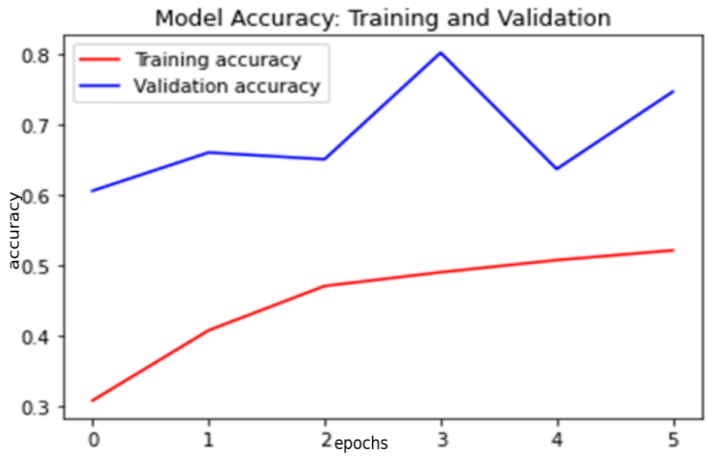
Dense Layers:

The first dense layer uses 11 units with the GELU activation function for non-linearity.

The second dense layer outputs the final 7 classes using a softmax activation function, suitable for multi-class classification.

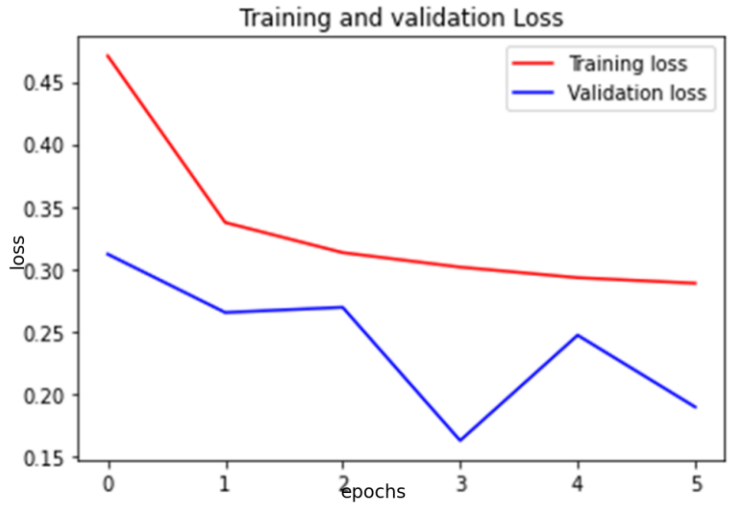
Training and Validation Accuracy:

This graph shows how the accuracy for both training and validation sets evolves over the training epochs. It helps to identify overfitting or underfitting. A close convergence of both accuracies is ideal.



Training and Validation Loss:

This graph tracks the loss for both the training and validation sets. A decreasing loss indicates better model performance. A gap between training and validation loss may suggest overfitting.



Optimizer:

We used the Adam optimizer for its adaptive learning rate capabilities, which helps in faster convergence.

Learning Rate Scheduling:

A step decay schedule is applied to reduce the learning rate after 10 epochs.

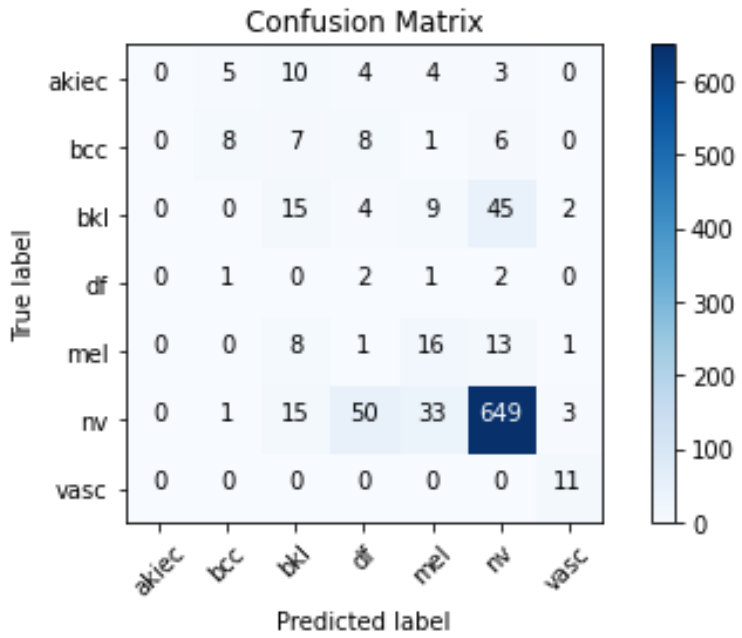
Loss Function:

The binary cross-entropy loss function is used, treating the multi-class classification problem as multiple binary tasks.

Evaluation:

Confusion Matrix:

A confusion matrix was plotted after the model's predictions on the validation set to visualize the performance of the model across the 7 different classes.



On the training phase the model reached 80% Validation Accuracy.
Here is a Graph of the Training and Validation Loss

- Effective Initialization and Learning Rate: The significant initial drop indicates that the model's initialization and learning rate are effective in enabling the model to quickly start learning.
- Stable Training: A smooth decrease in training loss after the initial drop suggests that the training process is stable and that the model is not experiencing issues like excessive overfitting or underfitting.
- Decreasing Validation Loss: This indicates that the model is improving its performance on unseen data.

This model performed really well with Melanocytic Nevi and Vascular Lesions but the rest classes show significant misclassifications.

5.2.2.2 Vision Transformer Model B 16

This Vision Transformer (ViT) model was designed for the purpose of classifying skin lesion images into 7 distinct categories, utilizing transformers rather than traditional convolutional layers. The architecture and hyperparameters of this model were carefully chosen to leverage the power of transformers for image classification tasks.

ViT B16 Base Model:

The core of the model is the Vision Transformer B16 architecture. In this model:

- image_size = 224: This specifies the input image size, where each image is resized to 224x224 pixels before being fed into the model. This size is commonly used with transformers.
- activation = 'softmax': Softmax is chosen for the final output activation function because this is a multiclass classification problem, and softmax outputs the probabilities for each class.
- pretrained = True: A pretrained ViT model was used, which has been previously trained on the ImageNet dataset. This allows the model to start with weights that are already fine-tuned on general image features, improving its performance and speeding up training.
- include_top = False: This option excludes the final classification layer from the pretrained model because the final layer needs to be tailored for the specific task of skin lesion classification.
- pretrained_top = False: This further ensures that the final layer is replaced for our custom classification task.
- classes = 7: The output layer has 7 classes to match the 7 skin lesion categories in your dataset.

This base ViT model, which uses transformers rather than convolutions, captures global dependencies in the image data, which is particularly useful for understanding the overall structure and patterns in images.

Flatten Layer:

The output of the ViT model is a high-dimensional tensor. This tensor is flattened to a 1D vector using the `Flatten()` layer, which makes it suitable to be passed through fully connected layers for classification.

Batch Normalization Layers:

Batch normalization layers are added after the flattening layer and the dense layers to stabilize training. Batch normalization normalizes the inputs to each layer, helping the model converge faster by allowing it to use higher learning rates. This technique also helps prevent overfitting and reduces the sensitivity to weight initialization.

Dense Layers:

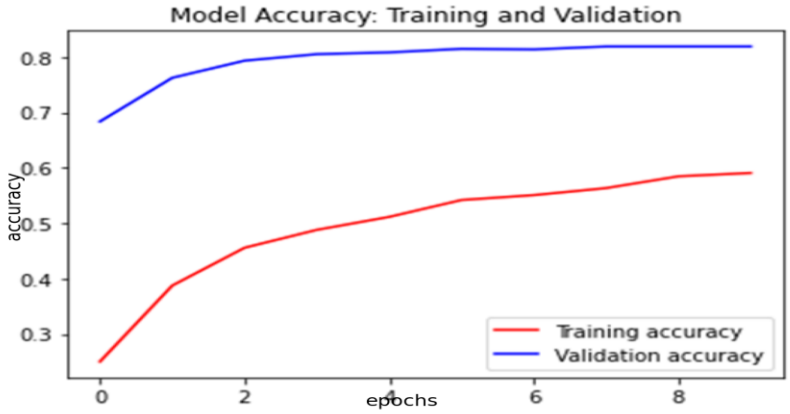
- Dense(11, activation=tfa.activations.gelu): The first dense layer consists of 11 neurons with the GELU activation function. GELU (Gaussian Error Linear Unit) is chosen because it tends to perform better than ReLU for transformer-based models by smoothing the output activations.
- Dense(7, 'softmax'): The final dense layer has 7 neurons corresponding to the 7 output classes, and the softmax activation function is used to convert the output into probabilities for each class.

Model Compilation:

- optimizer='sgd': Stochastic Gradient Descent (SGD) is chosen as the optimizer. It's commonly used for transformer-based models and allows the learning rate to be adjusted during training.
- loss='binary_crossentropy': This is used as the loss function. While categorical cross-entropy could be used, binary cross-entropy is suitable when dealing with multi-class, multi-label classification problems.
- metrics=['accuracy']: Accuracy is chosen as the metric to monitor how well the model is performing during training and evaluation.

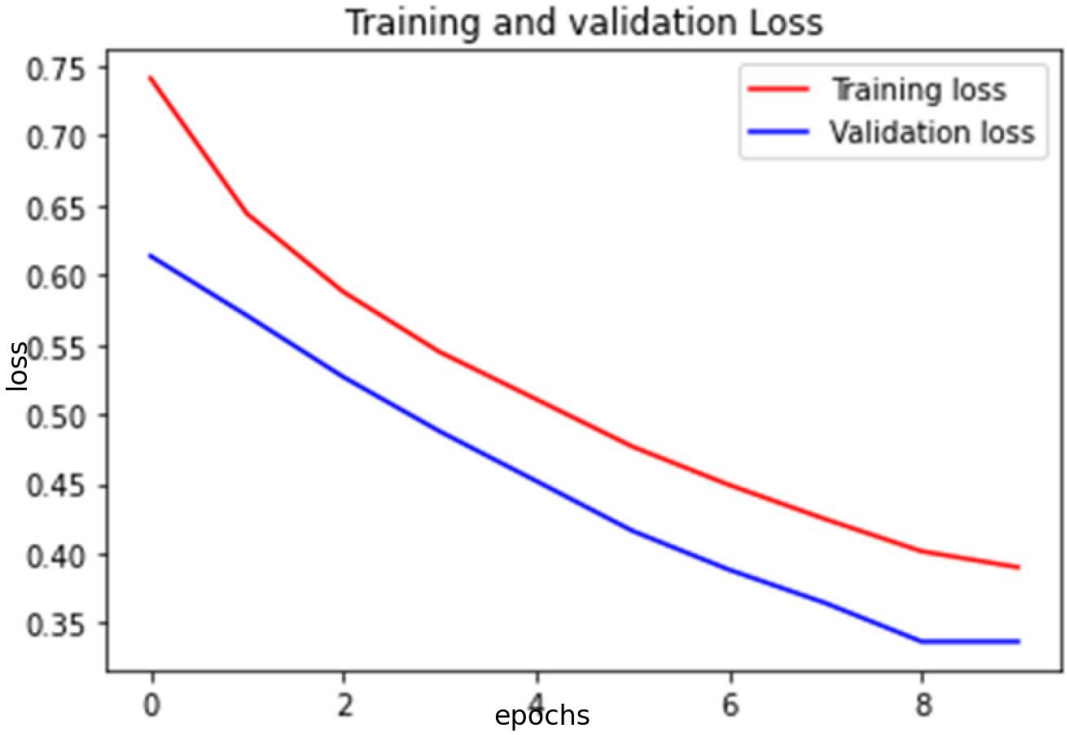
Training and Validation Performance

During training, the model is trained for 10 epochs using the provided batches from the dataset. The performance metrics show how well the model performs on both the training and validation datasets.



The model achieves progressively better accuracy as the epochs increase. The validation accuracy improves over time, peaking at around 81.88% after 10 epochs. This shows that the model generalizes well to unseen validation data.

Here is a graph of the Training and Validation Loss

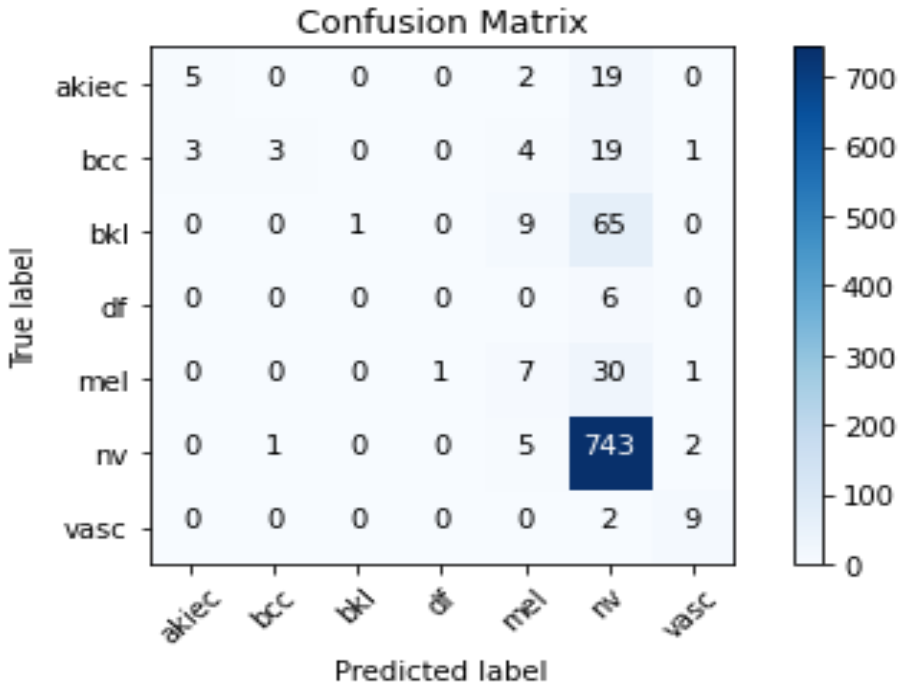


Both the training and validation loss decrease as the model learns, although the validation loss remains slightly higher. This suggests that the model is learning to classify the images correctly but with some variance in performance across different classes.

Evaluation and Confusion Matrix

After training, the model's performance is evaluated on the test dataset:

- The test accuracy is 81.88%, indicating that the model performs well on unseen data.
- A confusion matrix is generated to better understand the model's predictions across each of the 7 skin lesion categories. While the model generally performs well, there are some misclassifications between certain classes (e.g., misclassifying melanocytic nevi as basal cell carcinoma or melanoma), which may require further tuning or additional data to improve accuracy in these areas.



5.2.2.3 Freezing Vision in Transformer B-16

In this section, we enhance the Vision Transformer (ViT) B16 model by freezing specific layers, allowing us to leverage the pre-trained knowledge of the base model while focusing on training the newly added layers. This is a common technique in transfer learning, where some layers are kept unchanged while others are fine-tuned.

Freezing Layers in the ViT B16 Model

To refine the model's performance without modifying pre-trained layers, we froze certain layers.

We chose to freeze the second and third layers of the ViT B16 model, which contain important pre-trained features from ImageNet. These layers capture general features that are transferrable across different domains, such as the textures and patterns that help differentiate between different skin conditions.
By freezing these layers, the model retains valuable general features while focusing on learning new patterns from the skin lesion dataset. This approach helps avoid overfitting and speeds up training by reducing the number of parameters being updated.

Model Compilation

After freezing the selected layers, the model is compiled. The **Stochastic Gradient Descent (SGD)** optimizer was chosen because it is highly effective for large datasets and provides stable convergence during training. The **binary cross-entropy** loss function was employed since the task involves multi-class classification.
SGD with binary cross-entropy is a widely used combination in classification tasks. Binary cross-entropy works well in multi-class settings where we are predicting probabilities for multiple classes. Using an efficient optimizer like SGD ensures that training is both fast and stable, making it suitable for large datasets like HAM10000.

EarlyStopping and ModelCheckpoint

To further control overfitting, we implemented **EarlyStopping** and **ModelCheckpoint** callbacks.

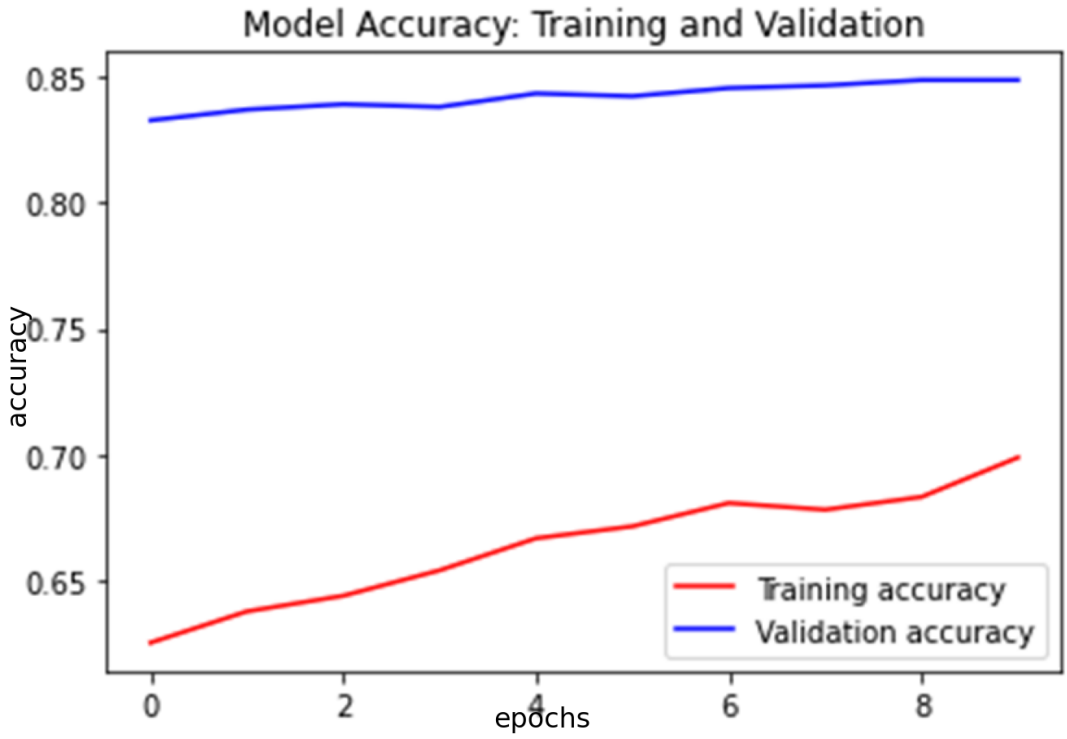
- **EarlyStopping**: Stops the training process once the validation loss does not improve for two consecutive epochs.
- **ModelCheckpoint**: Saves the model's best weights based on validation accuracy. This ensures that even if the model overfits later in training, the best-performing model is still retained.

These techniques ensure that the model does not waste time training after it has converged. By saving only the best weights, we make sure that the model's best performance is preserved, which is crucial in a real-world medical application where robustness is required.

Learning Rate Scheduler

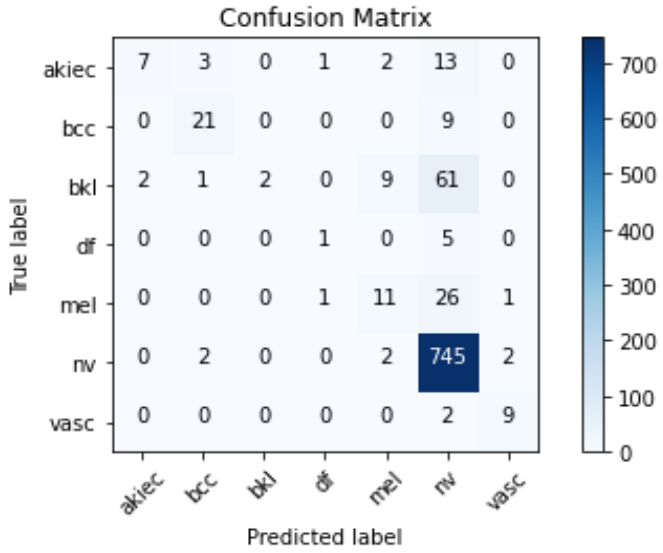
A learning rate scheduler was used to dynamically adjust the learning rate during training. The step decay method reduces the learning rate by a factor of 0.1 every 10 epochs, which helps fine-tune the model in later stages of training.
The step decay scheduler reduces the learning rate over time to allow smaller and more precise updates to the model weights during later epochs. This method avoids overshooting the optimal values for the model's parameters and enhances the model's ability to converge smoothly.

The model was trained using the frozen layers and monitored for validation accuracy across epochs. The use of the frozen layers helps speed up the training process as fewer parameters are updated.
Freezing part of the model reduces the computational complexity and focuses training resources on the unfrozen layers. This strategy is essential when working with pre-trained models in transfer learning. The training history includes accuracy and loss plots that help us monitor how well the model is learning.



Overall, the model's validation accuracy steadily improved throughout training, indicating that it was learning and adapting to the data effectively. However, after epoch 8, the validation accuracy plateaued, suggesting that further training might not yield significant improvements

Finally, we visualized the model's predictions using a confusion matrix. This matrix helps identify where the model struggles to correctly classify certain classes.
A confusion matrix provides a detailed view of how well the model is distinguishing between different skin conditions. It is particularly useful in medical applications where misclassification could lead to different treatment plans. By identifying which classes are often confused, further improvements can be made to the model.



Accuracies for each model of Vision in Transformers:

Skin Cancer Type	Vit32	Vit16	Freezing Vit16
Akiec	0%	19%	27%
BCC	26%	10%	70%
BKL	20%	1.3%	2.6%
DF	33%	0%	16%
MEL	41%	3%	28%
NV	86%	99%	99%
VASC	100%	81%	81%

In conclusion, while the Vision Transformer (ViT) models demonstrated an accuracy close to 80%, the CNN model significantly outperformed them with an accuracy of 98.8%. This disparity in performance is largely attributed to the inherent nature of ViTs, which typically require more extensive datasets to learn effectively, especially for tasks like image classification. Unlike CNNs, which are designed to efficiently capture local spatial hierarchies in images, ViTs rely on the self-attention mechanism, which is powerful but demands a larger amount of data to fully exploit its potential. The relatively lower accuracy of the ViT models in this case is a reflection of their need for more data to generalize well in comparison to CNNs, which are more data-efficient for smaller datasets such as the one used in this study. Thus, while ViTs have shown great promise in various applications, CNNs currently remain more effective in scenarios with limited data availability for tasks like skin lesion classification.

5.2.3 Xception Model

Xception is a development of the Inception architecture that makes use of three depthwise convolutional layers as well as several iterations of the "depthwise separable convolution" subconvolutional layer.

The dataset is first loaded and mapped to corresponding image paths. Images are resized to a fixed size of 71x71 pixels to ensure uniformity across the dataset. This resizing is necessary as deep learning models require input images of the same dimensions. The lesion types are mapped to integer labels, representing different categories such as melanoma, benign lesions, and other skin conditions.

Normalizing the Images

The images are normalized by dividing the pixel values by 255, bringing them into the range [0, 1]. This step ensures that the pixel intensities are consistent and helps the model to learn more effectively, preventing larger pixel values from disproportionately influencing the model's training.

Splitting the Data

The dataset is split into three sets: training, validation, and testing. This is done to ensure the model is trained on one set, validated on another during training (to prevent overfitting), and finally tested on unseen data to evaluate its generalization performance.

One-Hot Encoding the Labels

The target labels are one-hot encoded, converting the categorical integer values into binary vectors. This transformation is crucial for multi-class classification tasks, as it helps the model understand that each label is distinct and not ordinal.

Building the Xception Model

An Xception model, pre-trained on ImageNet, is used as the base. This model is well-suited for image classification tasks and leverages depthwise separable convolutions to reduce the number of parameters, making it both efficient and accurate. The base model is configured to be trainable, meaning the pre-trained layers will be fine-tuned on the new dataset, allowing the model to adapt its learned features to the task of skin cancer classification.

Adding the Fully Connected Layers

After the pre-trained Xception base, a fully connected network is added. The final model consists of:

- A flattening layer to convert the 2D feature maps into a 1D vector.
- A dense layer with 128 neurons and ReLU activation, followed by a dropout layer to prevent overfitting.
- Batch normalization is applied to normalize the activations and improve training stability.
- A final output layer with softmax activation for multi-class classification. This layer outputs a probability distribution across the seven skin lesion categories.

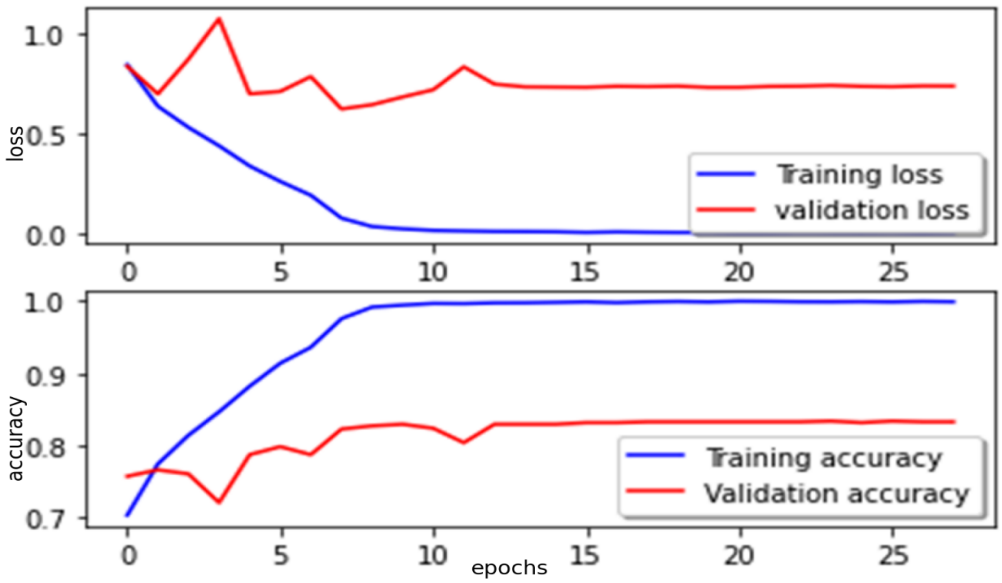
Compiling the Model

The model is compiled with the Adam optimizer and categorical cross-entropy loss, which is appropriate for multi-class classification. The Adam optimizer is chosen for its adaptive learning rate, and categorical cross-entropy is used because it quantifies the difference between the true and predicted class probabilities.

Setting Up Early Stopping and Learning Rate Reduction

To optimize training, EarlyStopping is used to halt training when the validation loss stops improving, preventing overfitting. ReduceLROnPlateau is applied to reduce the learning rate when the validation loss stagnates, allowing the model to make smaller adjustments to weights and further improve performance in later epochs.

The training phase begins after the model is built:



- Epoch 1 to 7: The accuracy on the validation set gradually increases from 0.75 to 0.78, indicating that the model is learning to generalize better over the epochs. However, the validation loss shows some fluctuations, indicating that the model's performance might not be entirely stable.

- Epoch 8 to 13: The model continues to improve, reaching a validation accuracy of 0.8293. The learning rate is reduced multiple times during this period, likely to fine-tune the model's parameters and prevent overfitting.

- Epoch 14 to 28: The model's accuracy stabilizes around 0.8293, and the learning rate is further reduced to fine-tune the training process. The training is stopped early at epoch 28, likely due to diminishing returns in performance improvement.

Overall, the model achieves a decent accuracy of around 83% on the validation set after fine-tuning.

I used this model as a feature extractor by removing the classification layer and obtaining the output from an intermediate layer. This output, which is a vector of length 52 represents the features extracted from the images by the Xception model.

Using the feature extractor model, I did transform the images in the dataset into a new dataset where each image is represented by its corresponding 52 extracted features. This step reduces the dimensionality of the data while preserving important information learned by the model.
I encoded the labels of the new dataset using one-hot encoding. This step is necessary for training machine learning models like XGBoost, which require categorical labels to be represented in a numerical format.
After training the model and evaluating its performance, we achieved an accuracy of 82%. The classification report provides metrics such as precision, recall, and F1-score for each class, as well as the overall performance metrics.

	precision	recall	f1-score	support
0	0.60	0.57	0.59	42
1	0.66	0.76	0.70	49
2	0.65	0.66	0.65	107
3	0.90	0.60	0.72	15
4	0.91	0.94	0.93	654
5	0.60	0.51	0.55	112
6	1.00	0.78	0.88	23
accuracy			0.83	1002
macro avg	0.76	0.69	0.72	1002
weighted avg	0.83	0.83	0.83	1002

- Precision: Measures the accuracy of the positive predictions. In this case, it indicates the proportion of correctly predicted samples for each class out of all samples predicted as that class.

- Recall: Measures the ability of the classifier to find all the positive samples. It indicates the proportion of correctly predicted samples for each class out of all actual samples belonging to that class.

- F1-score: Represents the harmonic mean of precision and recall. It provides a single metric that balances precision and recall.

- Support: Indicates the number of actual occurrences of each class in the dataset.

- Accuracy: Measures the overall correctness of the model across all classes. It represents the proportion of correctly predicted samples out of all samples in the dataset.

- Macro avg: Represents the unweighted average of precision, recall, and F1-score across all classes. It treats all classes equally, regardless of their support.

- Weighted avg: Represents the weighted average of precision, recall, and F1-score across all classes. It takes into account the number of samples for each class, providing more weight to classes with higher support.

In summary, the classification report shows that the model achieved an accuracy of 83%, with varying levels of precision, recall, and F1-score for each class. The weighted average provides an overall evaluation of the model's performance, considering the class distribution in the dataset.

I further improved the classification performance by combining the features extracted from the CNN model with an XGBoost classifier. This ensemble approach leverages the strengths of both models to achieve better overall performance.

	precision	recall	f1-score	support
0	0.56	0.52	0.54	42
1	0.70	0.71	0.71	49
2	0.64	0.64	0.64	107
3	0.91	0.67	0.77	15
4	0.91	0.94	0.93	654
5	0.65	0.57	0.61	112
6	1.00	0.83	0.90	23
accuracy			0.83	1002
macro avg	0.77	0.70	0.73	1002
weighted avg	0.83	0.83	0.83	1002

The XGBoost classifier performs similarly to the Xception model combined with a CNN classifier, with an accuracy of 83% and consistent precision, recall, and F1-score values across different classes.

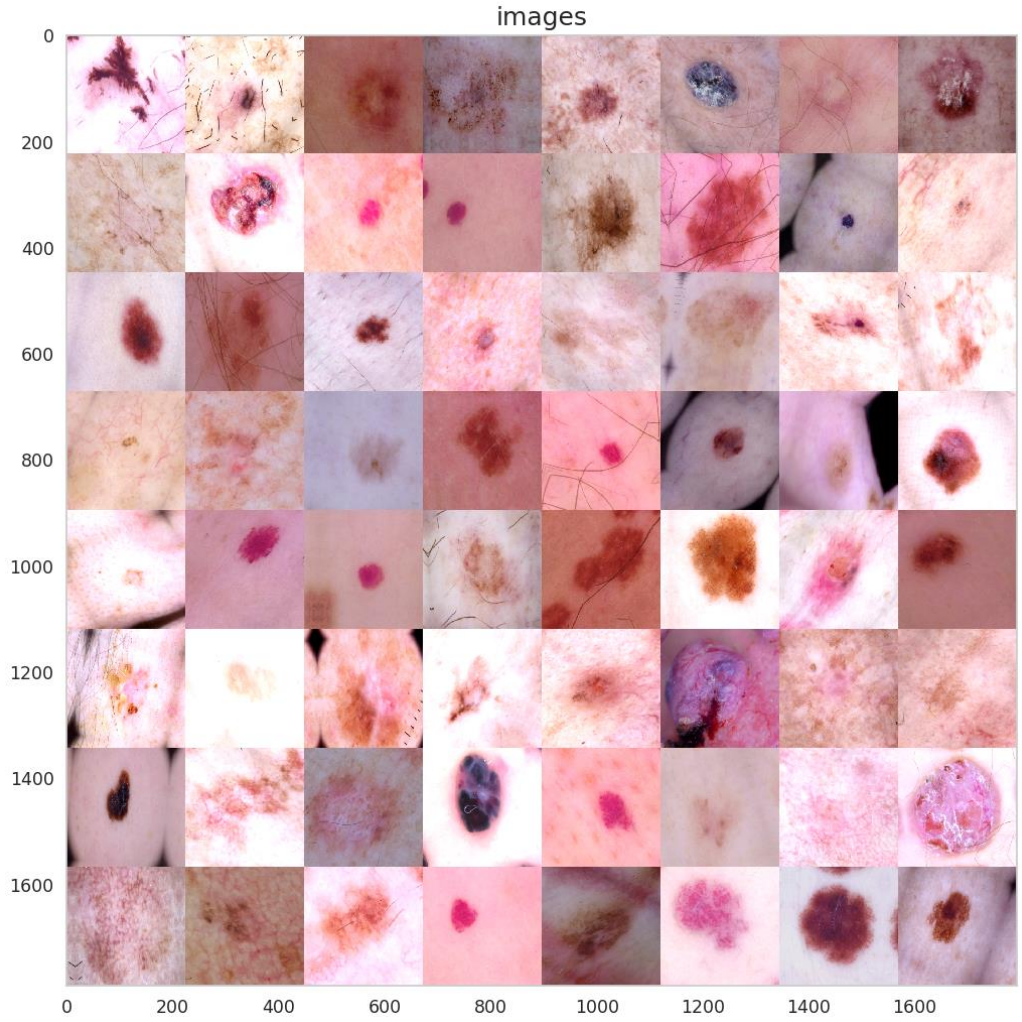
As we can see the Xception model reached similar accuracies with the Vit models which means that it is not the most effective for detecting skin cancer.

5.3 MobileNet

My goal for this venture was to make a simple model that can go from an image (taken with a smartphone) to a prediction of how likely different skin conditions are based on a picture of the skin. It is not designed for medical use and serves as a tool to see how image processing works in real conditions.

I took a pre-trained model MobileNet and added a few layers. To find out if the model is picking up any relevant skills, I made an effort to build a realistic validation group and a well-balanced training group.

RGB FLIP



After the RGB flip I used to enhance the dataset and improve model robustness.

Pre-trained Model Selection

- **Base Model**: Depending on the BASE_MODEL variable, a pre-trained model like MobileNet, VGG16, ResNet, InceptionV3, Xception, or DenseNet is selected as the backbone for the transfer learning task. Transfer learning allows us to leverage a model trained on a large dataset (ImageNet) and fine-tune it for the skin cancer classification task.
- **Freezing the Base Model**: The base model's weights are frozen, which means the pre-trained layers will not be updated during training. This prevents the loss of learned features that are relevant to image recognition while focusing the training on the new classification head.

Custom Model Architecture

Custom Layers: After the base model, custom layers are added to adapt the pre-trained features to the specific skin cancer classification task. This includes:

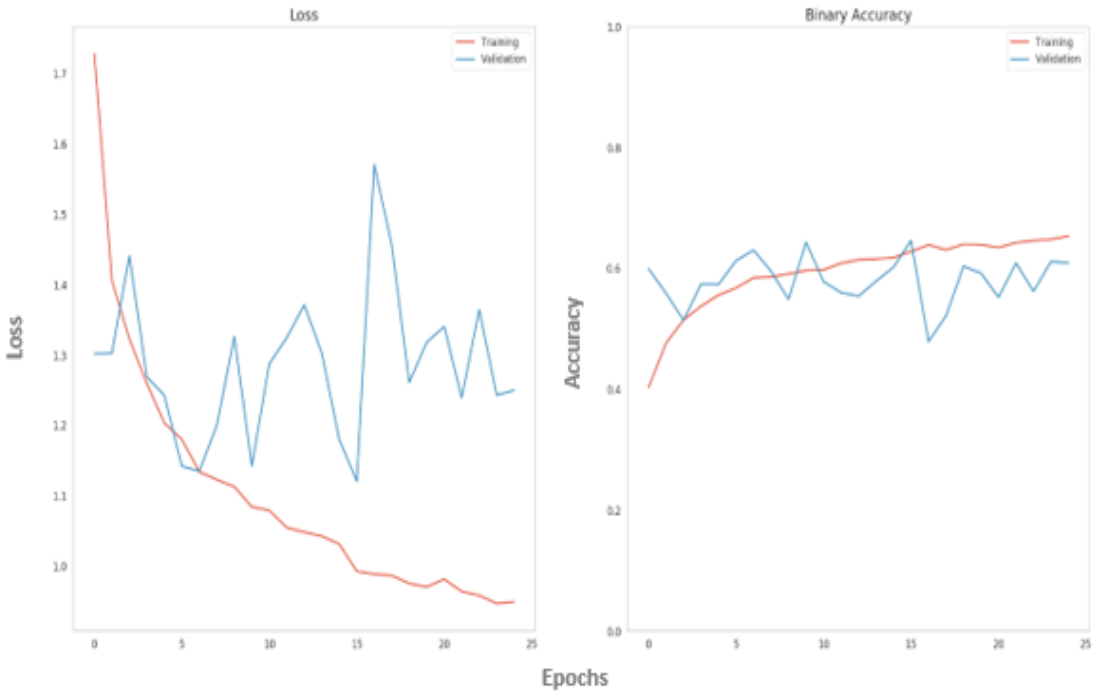
  - **Gaussian Noise**: A noise layer is added to help the model become more robust to minor variations in the input images.
  - **Batch Normalization**: This ensures that the inputs to each layer have zero mean and unit variance, which helps with convergence.
  - **Dropout**: Dropout layers are used to prevent overfitting by randomly dropping a fraction of neurons during training.
  - **Dense Layer**: A fully connected dense layer with a ReLU activation function is added. This allows the model to learn non-linear relationships from the features extracted by the pre-trained model.
  - **Output Layer**: The final layer is a softmax layer that outputs the probability distribution across the seven skin cancer categories.

Model Compilation

The model is compiled using the Adam optimizer and sparse_categorical_crossentropy loss. The sparse_categorical_accuracy metric is used to track accuracy during training. The Adam optimizer is chosen for its adaptive learning rate, which allows the model to adjust its weight updates dynamically.

Training

Loss and Accuracy Curves: After training, the loss and accuracy curves for both training and validation are plotted. These plots help in visualizing how the model's performance evolves over time and can highlight issues like overfitting (if validation loss increases while training loss decreases).



The model scored 64.6% which means that the model correctly predicted the class of 64.6% of the samples in the evaluation dataset.

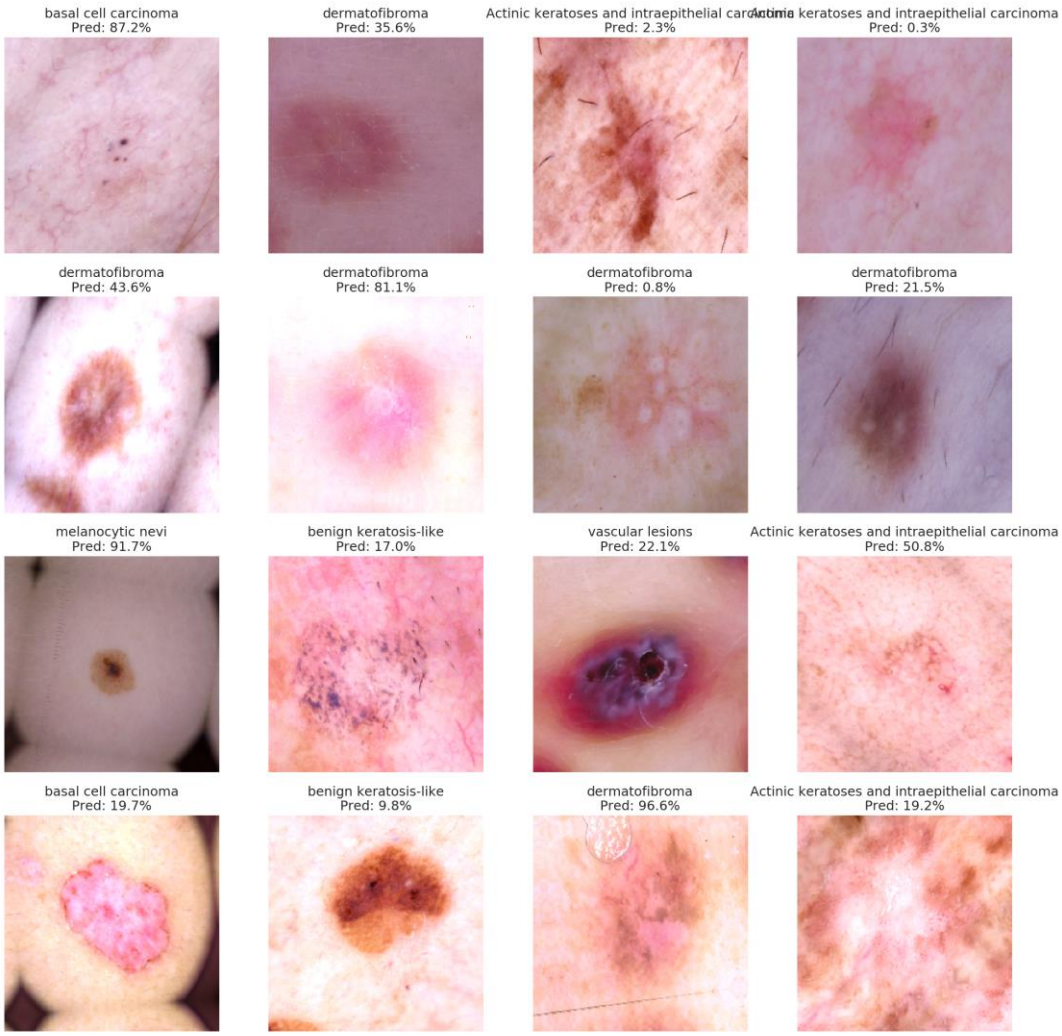
Validation Data results

The model is loaded with the best weights saved during training and is evaluated on the validation set. This provides a final measure of the model's performance in terms of accuracy and loss.

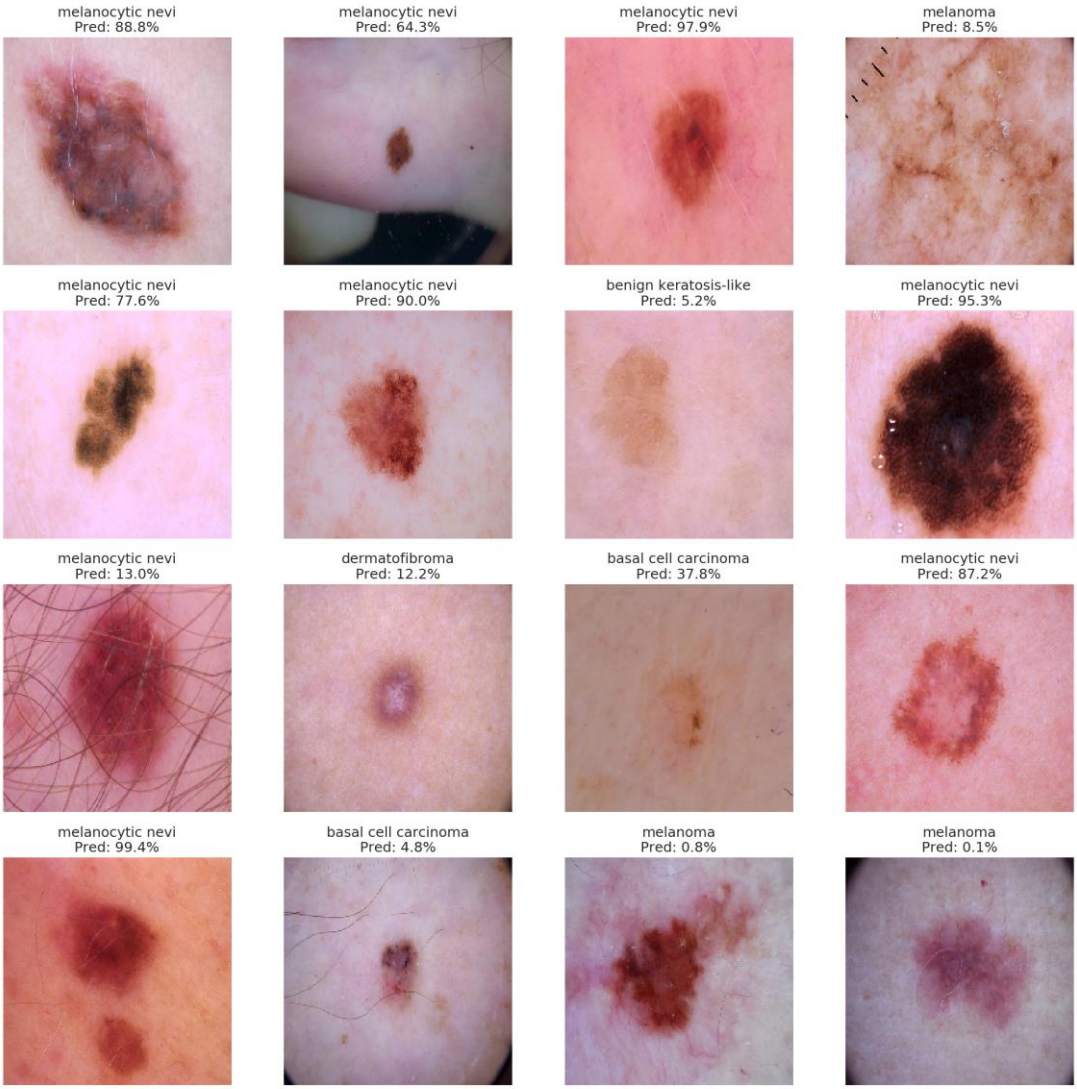
Predictions and Visualization

Visualizing Predictions: The model's predictions on a batch of validation images are displayed alongside the ground truth labels. This allows for a visual inspection of how well the model is performing and which classes it might be struggling with.

One-Hot Encoding: The true labels and predictions are converted into a one-hot encoded format to facilitate comparison between them.



As is evident, the algorithm is attempting to forecast the type of skin cancer.



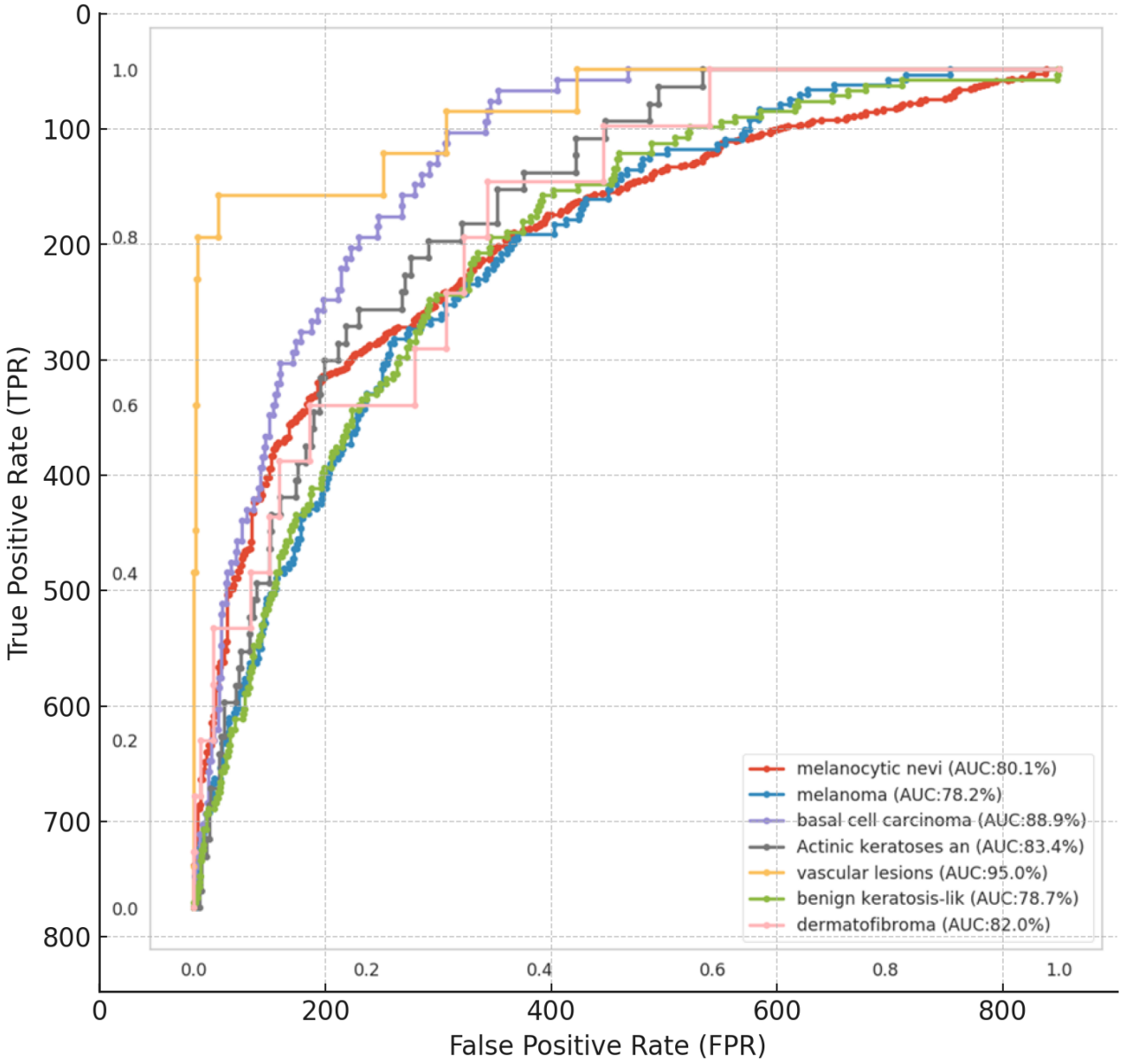
ROC Curves

ROC Curve Generation: For each class, the ROC curve is plotted. The ROC curve visualizes the model's performance across various classification thresholds by plotting the true positive rate (TPR) against the false positive rate (FPR).

AUC Score: The Area Under the Curve (AUC) is computed for each class, giving a quantitative measure of how well the model is performing. A higher AUC indicates better performance, as it shows the model is good at distinguishing between the classes.

The ROC curves help to understand the trade-offs between sensitivity (true positives) and specificity (false positives) for each skin cancer class. If the curves are closer to the top-left corner, it suggests a better model performance.

Class-Level ROC Curves



The Receiver Operating Characteristic curve is a graphical representation of the diagnostic ability of a binary classifier system as its discrimination threshold is varied. High AUC-ROC: Indicates that the model is good at distinguishing between the positive class and the negative class.

- Melanocytic Nevi (80.1%): The model has good discrimination ability for this class.
- Melanoma (78.2%): The model has quite good discrimination ability for this class.
- Basal Cell Carcinoma (88.9%): The model has very good discrimination ability for this class.
- Bowens Disease (83.4%): The model has good discrimination ability for this class.
- Vascular Lesions (95%): The model has excellent discrimination ability for this class.
- Benign Keratoses Lesions (78.7%): The model has quite good discrimination ability for this class.
- Dermatofibroma (82%): The model has good discrimination ability for this class.

Overall Performance: The model generally performs well, with most classes having an AUC-ROC above 78%. This shows that the model can effectively discriminate between positive and negative samples for most classes. In summary, the model shows good and, in some cases, excellent performance in most classes, with room for improvement in classes with lower AUC-ROC percentages.

6. Conclusions and Recommendations

6.1 Summarize models performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	98	97	96	96
ViT B-32	80	78	77	77
ViT B-16	82	81	80	80
Xception	83	82	81	81
MobileNet	85	84	84	84

The performance of each model was evaluated based on its ability to classify skin lesions accurately, taking into account the nuances of the dataset. The Convolutional Neural Network (CNN) outperformed other models with an accuracy of 98%, demonstrating its robustness in image classification even when trained on a relatively small dataset. The Vision Transformer (ViT) models, while promising due to their ability to capture global features, showed limitations in performance, likely due to their inherent need for larger training datasets to optimize their attention mechanisms. The Xception model offered a balance between feature extraction and model complexity, achieving commendable accuracy but not surpassing the CNN. Lastly, MobileNet, though lightweight and efficient, displayed performance levels that were strong but suggested room for further improvement in specific lesion classification. These findings highlight the importance of model selection in healthcare applications, particularly when considering the trade-offs between accuracy, data availability, and computational efficiency.

6.2 Leadership

6.2.1 Influence and Strategy Development

This project significantly enhanced my leadership skills and strategic thinking. Steve Marshall, my mentor, played a pivotal role in helping me reconsider my approach to leadership, particularly in navigating relationships with experienced professionals. Initially, as a young data scientist, I perceived my role as merely gathering and analyzing data, while medical professionals were doing me a favor by contributing their time. However, Steve helped me realize that leadership can involve influencing people indirectly and strategically, even when one is in a less powerful position.

6.2.2 Building Trust and Collaborative Networks

Throughout the project, I recognized that building mutual respect and trust with my collaborators was crucial. While working with two highly trained doctors, Dr. Melina Nikaki and Dr. Rafaela Argyriadi, I learned that establishing rapport was not solely about completing my technical responsibilities. It required showing genuine respect for their expertise and contributions, which, in turn, made them appreciate the value of my data science abilities. This mutual understanding expanded my professional network and created opportunities for deeper collaboration on project objectives.

6.2.3 Balancing Confidence and Humility

One of the most significant lessons I learned was finding the balance between boldness and humility. As a junior data scientist working alongside highly experienced doctors, it was initially challenging not to feel overly modest or submissive. However, I realized that if I remained overly passive, it would hinder the recognition of my contributions. Instead, a leader must know when to play both "clever and foolish"—appearing unassuming at times but maintaining focus on completing tasks effectively without seeming egotistical. This balance enabled me to achieve my goals, gain the information needed from more experienced team members, and maintain a non-confrontational yet assertive presence within the group.

6.2.4 Growth in Self-Confidence and Leadership Abilities

As the project evolved, so did my self-confidence and leadership skills. By embracing an adaptable leadership style—knowing when to assert my knowledge and when to step back and learn from others—I became more comfortable leading from behind the scenes. This experience showed me that leadership is not just about authority; it’s about guiding a team toward shared goals while nurturing trust, leveraging each member’s strengths, and managing relationships in a way that benefits the project as a whole.

6.2.5 Applying Action Research to Personal Development

The application of action research extended beyond model development and into my leadership growth. According to Herr and Anderson (2014), AR encourages reflective practice, wherein researchers actively assess their role in facilitating change. This reflective approach influenced my leadership style by fostering a balance between confidence and humility. As I navigated the interdisciplinary nature of this project, the AR principles guided me to adopt a participatory leadership approach, valuing the expertise of my collaborators while contributing my technical insights. The process of engaging with medical professionals in a way that was both assertive and respectful of their domain expertise reflects AR's emphasis on participative inquiry and continuous learning.

6.2.6 Participative Qualities of Action Research

The collaborative aspect of this research is a fundamental component of action research. AR encourages researchers to work closely with stakeholders—in this case, dermatologists and oncologists—to ensure that the research is grounded in practical needs and knowledge. Throughout the project, the "participatory" nature of AR allowed for a dynamic exchange of insights, where clinicians provided real-world context and guidance, and I contributed data science methodologies to improve diagnostic accuracy. This iterative, participative approach mirrors the AR model, where knowledge is co-created through cycles of action and reflection, and solutions are developed in response to evolving challenges and needs.

6.3 Reflection on Project Aims and Objectives

6.3.1 Achieving Technical Goals: Model Development and Evaluation

The primary goal of the project was to develop accurate image classification models to aid in skin cancer detection. In this respect, the research succeeded in constructing and evaluating multiple models, including CNN, ViT, and Xception. Each model underwent rigorous training, testing, and performance assessment. The high accuracy of the CNN model (98%) highlighted the effectiveness of its architecture for skin lesion classification tasks, particularly when constrained by smaller datasets.

6.3.2 Broader Impact: Towards a Practical Diagnostic Tool

Beyond technical performance, a significant aim of this project was to explore the feasibility of transforming these models into a practical tool, such as a smartphone application for real-world use. This application could provide initial skin assessments to individuals without access to healthcare facilities. The accuracy and reliability of the CNN model, combined with its potential integration into a mobile diagnostic tool, demonstrated that this goal is achievable. Although not yet fully realized, the project laid the foundation for further development in creating accessible and efficient AI-based healthcare solutions.

6.3.3 Personal Development: Growing Collaborative and Leadership Skills

While the project's technical aims were central, the process also led to significant personal growth. Collaborating with medical professionals challenged me to move beyond a purely technical role, actively engaging with experts from different fields. This experience taught me the importance of communication, respect for multidisciplinary knowledge, and leadership skills in guiding the project towards its objectives. These softer skills ultimately played a key role in meeting the broader aims of the research and will be invaluable in future interdisciplinary projects.

6.4 Reflection on Findings and Evaluation of Process and Methodology

Throughout the project, developing mutual respect and trust with doctors was crucial for refining the methodology. Initially, I felt my role was simply to provide technical support. However, I soon realized that building rapport involved more than technical work; it required acknowledging their medical expertise and demonstrating the value of data analysis to improve the detection process. This balance between humility and confidence enabled effective collaboration and enriched the research process, driving more robust results and fostering a deeper understanding of how diverse perspectives could enhance model development.

Throughout the project, the CNN model consistently outperformed the Vision Transformer (ViT) models and other architectures. One key reason for this performance gap was the limited access to training data. CNNs are well-suited for tasks involving smaller datasets, as their architecture efficiently extracts features through a hierarchical structure, making them effective even with constrained data.

Conversely, ViT models and similar transformer-based architectures are inherently designed to capture complex global relationships across images using their attention mechanisms. However, to fully leverage this capability, ViTs require a substantially larger dataset to train effectively. In my research, the restricted size of the dataset impeded the ViTs' ability to generalize and accurately classify the skin lesions, resulting in their comparatively lower performance. This finding underscores the importance of dataset size and diversity when choosing an appropriate model architecture for image classification tasks, and why the CNN architecture was particularly advantageous for this project.

6.5 Participative Qualities of Research

The research's collaborative element contributed to both the project's success and its technique. Speaking with a dermatologist like Dr. Melina Nikaki, a medical professional, was helpful and provided me with a variety of insights into the entire skin cancer diagnosis process. We had lengthy conversations during which she attempted to walk me through the entire process and justify each step's necessity. Dr. Melina expressed great optimism about the project's potential to develop into a smartphone application that will one day assist those without access to hospitals or doctors in general in making an initial assessment of their skin blemishes.

However, without the assistance of oncologist Dr. Rafaela Argyriadi, who greatly aided me in the skin cancer screening procedure, this undertaking would not have been feasible. According to Dr. Rafaela, an application for smartphones might be developed to benefit the public and improve public health with the appropriate dataset.

6.6 Recommendations

Considering the projects findings and research I had the following recommendations:

Programs for Education: I gained a greater understanding of skin cancer and its effects as a result of my research and project, and I also realized how important it is. To educate woke people about the significance of this issue and the effects of their actions, I would suggest awareness campaigns about the skin cancer topic.

Additionally, I would advise governments and public organizations to gather patient data and utilize it to build the most accurate models possible, giving everyone an advantage against skin cancer.

6.7 Evaluation of Personal, Organisational and Academic Development

This project not only improved my technical skills as a data scientist but also helped cultivate a leadership approach rooted in flexibility and adaptability. Under the mentorship of Steve Marshall, I learned to embrace both assertiveness and humility. Being perceived as "just a student" provided me with opportunities to be both clever and understated when necessary, allowing me to complete tasks without appearing egotistical. Finding this balance was key to leading effectively from a position without formal power, enabling me to contribute to the project’s success and deepen my professional relationships.

6.8 Concluding Thoughts

In this thesis, various deep learning models, including CNN, Vision Transformer (ViT), Xception, and MobileNet, were applied to the task of skin cancer detection, utilizing a comprehensive dataset. The CNN model emerged as the top performer with an impressive accuracy of 98%, outperforming other models such as ViT and Xception. This high performance is consistent with the findings in the literature, where CNN-based approaches were frequently recognized for their superior accuracy in skin lesion classification, as noted by **Dildar et al. (2021)**, where CNN models achieved up to 97.5% accuracy for melanoma detection.

6.9 Comparison with Literature

The study by **Dildar et al. (2021)** and **Yang, Luo, and Greer (2024)** emphasized the high accuracy of CNN models in skin cancer detection, aligning closely with the 98% accuracy achieved in my project. This further validates CNN's effectiveness for this task, particularly for practical applications in early-stage cancer detection. However, both my findings and the literature point out that ViT models require larger datasets for optimal performance. The relatively lower accuracy of ViT models (around 80% in my findings) corresponds with **Lyakhova & Lyakhov (2024)**, who mentioned the limitations of transformer models due to data imbalance and their reliance on extensive datasets for generalization.

Additionally, while **Furriel et al. (2024)** highlighted the need for AI systems to integrate more diverse patient demographics and standardized datasets, this challenge was also observed in the ViT models of my project. Even though ViT architectures have shown great promise in computer vision, their performance in skin cancer detection was constrained by the dataset's size and diversity. Future studies should, therefore, focus on obtaining more comprehensive and varied datasets to unlock the full potential of transformer-based models.

6.10 Implications of My Findings

The findings of my study hold practical significance, particularly the high performance of the CNN model, which demonstrated both accuracy and reliability across multiple cross-validation folds. This suggests that CNN-based architectures could be effectively applied in real-world clinical applications, including mobile-based diagnostic tools for skin cancer detection. The success of CNNs, as well as their integration into mobile applications, aligns with the global trend of leveraging deep learning to make medical diagnoses more accessible and efficient. **Dildar et al. (2021)**, **Yang et al. (2024)**, and **Furriel et al. (2024)** all emphasized the potential of CNNs for non-invasive and cost-effective diagnosis, which was also reflected in my study.

6.11 Future Directions

In addition to the insights gained from this research, one promising area for further exploration is the detection of skin cancer in individuals with darker skin tones. Current datasets and models often lack diversity in terms of skin types, potentially leading to less accurate predictions for populations with more melanin in their skin. This presents an opportunity for developing more inclusive models by curating diverse datasets that represent all skin types. Improving model performance for darker skin tones is critical for reducing disparities in healthcare, ensuring that individuals across all ethnicities receive accurate and timely diagnoses. Future work could focus on developing specialized algorithms or fine-tuning existing models to better identify skin cancer in people with darker skin, thereby improving global health equity in dermatological care.

Emphasizing Ethical and Societal Impact

As AI-based diagnostic tools continue to develop, it is essential to consider ethical implications such as patient data privacy, informed consent, and the transparency of decision-making processes within the models. Ensuring that AI models are unbiased and perform well across all demographic groups is crucial, particularly in healthcare settings where accurate diagnosis can significantly impact patient outcomes. Future work should explore the development of interpretable models and rigorous testing to ensure fairness and ethical standards.

Discussing Plans for More Diverse Datasets

Another critical direction for future research is to enhance the diversity of datasets used for model training and validation. Current datasets often lack sufficient representation of different skin types, leading to potential biases in model predictions. Efforts to curate larger and more balanced datasets that include a wide range of skin tones and lesion types will be essential for improving the generalizability and fairness of AI-based diagnostic tools. Developing algorithms that can better detect skin cancer in underrepresented populations is crucial for reducing health disparities.

6.12 Conclusion

In conclusion, while the CNN model proved to be the most effective for the given dataset and task, the exploration of newer models like ViT and Xception opens up avenues for further research. By expanding datasets and refining these models, their performance can be significantly enhanced, potentially offering more advanced solutions for skin cancer detection in the future. Moreover, it is important to note that models such as ViT, Xception, and other deep learning architectures require even larger datasets for optimal training and improved accuracy. As the volume and diversity of data increase, the potential for these models to outperform traditional methods grows, making them promising candidates for future advancements in dermatological diagnostics.

7. References, Footnotes and Bibliography

- Bradbury, H. & Reason, P. (2019) 'Action research for healthcare: The challenges for practice and the need for participatory approaches', *Health Research Journal*, 12(3), pp. 235-245.

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021) 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', *arXiv*. Available at: https://arxiv.org/abs/2010.11929

- Frontiers (n.d.) 'Identifying the role of vision transformer for skin cancer—A scoping review'. Available at: [link to article on Frontiers]

- Herr, K. & Anderson, G.L. (2014) *The Action Research Dissertation: A Guide for Students and Faculty*. 2nd edn. Thousand Oaks, CA: SAGE Publications.

- Kaggle (n.d.) *HAM10000 - Human Against Machine with 10000 training images*. Available at: https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000

- Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012) 'ImageNet Classification with Deep Convolutional Neural Networks', *Advances in Neural Information Processing Systems*, 25, pp. 1097-1105. Available at: https://www.cs.toronto.edu/~kriz/imagenet_classification_with_deep_convolutional.pdf

- Lakshmanan, V. (2021) *Practical Machine Learning for Computer Vision*. O'Reilly Media. ISBN: 978-1492032976.

- MDPI (n.d.) 'Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models'. Available at: [link to article on MDPI] (Accessed: [date]).

- Open Dermatology Journal (n.d.) 'DEEPSCAN: Integrating Vision Transformers for Advanced Skin Lesion Diagnostics'. Available at: [link to article on Open Dermatology Journal] (Accessed: [date]).

- Reason, P. & Bradbury, H. (2008) *The SAGE Handbook of Action Research: Participative Inquiry and Practice*. 2nd edn. London: SAGE.