**Skin Disease Detection using Image Data**

***Project report submitted to***

***Indian Institute of Information Technology, Nagpur, in partial fulfilment of the requirements for the award of the degree of***

**Bachelor of Technology**

**In**

**Computer Science and Engineering**

***By***

|  |  |  |
| --- | --- | --- |
| **Aayush Patil (BT21CSE179)** | **Ankit Doye (BT21CSE170)** | **Vaibhav Taywade (BT21CSE200)** |

Under the guidance of

**Dr. Khushboo Jain**

****

**Indian Institute of Information Technology, Nagpur 441108 (India)**

**Batch 2025**

**© *Indian Institute of Information Technology, Nagpur* (IIIT) 2024**

Department of Computer Science Engineering

Indian Institute of Information Technology, Nagpur

**Declaration**

We, Aayush, Ankit and Vaibhav, hereby declare that this project work titled “Skin Disease Detection using Image Data” is carried out by us in the Department of Computer Science Engineering of Indian Institute of Information Technology, Nagpur. The work is original and has not been submitted earlier whole or in part for the award of any degree/diploma at this or any other Institution /University.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Enrollment No.** | **Name** | **Signature** |
| 1 | BT21CSE179 | Aayush Patil |  |
| 2 | BT21CSE170 | Ankit Doye |  |
| 3 | BT21CSE200 | Vaibhav Taywade |  |

**Declaration**

We, Aayush(BT21CSE179), Ankit (BT21CSE170) and Vaibhav (BT21CSE200) understand that plagiarism is defined as any one or the combination of the following:

1. Uncredited verbatim copying of individual sentences, paragraphs or illustration (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.

2. Uncredited improper paraphrasing of pages or paragraphs (changing a few words or phrases, or rearranging the original sentence order).

3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who did or wrote what. (Source: IEEE, the institute, Dec.2004) I have made sure that all the ideas, expressions, graphs, diagrams, etc. that are not a result of my own work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.

We affirm that no portion of my work can be considered as plagiarism and I take full responsibility if such a complaint occurs. I understand fully well that the guide of the thesis may not be in a position to check for the possibility of such incidences of plagiarism in this body of work.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Enrolment No.** | **Name** | **Signature** |
| 1 | BT21CSE179 | Aayush Patil |  |
| 2 | BT21CSE170 | Ankit Doye |  |
| 3 | BT21CSE200 | Vaibhav Taywade |  |

**Department of Computer Science Engineering**

**IIIT, Nagpur**

Date:

**Certificate**

This is to certify that the project titled “**Skin Disease Detection using Image Data**”, submitted by **Aayush, Ankit and Vaibhav** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science Engineering**, IIIT Nagpur. The work is comprehensive, complete and fit for final evaluation.

Date:

**Dr. Nishat A. Ansari**  **Dr. Khushboo Jain**

**Designation, Dept, IIIT, Nagpur Designation, Dept, IIIT, Nagpur**

**Acknowledgement**

It is with great pleasure that we present the report of our Project Work undertaken during the final year of our B.Tech. We would like to express our heartfelt gratitude to our Project Mentor, Dr. Jitendra Temburne, Department of Computer Science Engineering, Indian Institute of Information Technology, Nagpur, for his continuous support and invaluable guidance throughout the course of this project. His dedicated efforts have been instrumental in bringing our work to fruition, and words cannot fully convey our deep appreciation for his mentorship.

We would also like to extend our sincere thanks to the Department of Computer Science at IIIT Nagpur for their support and for providing the essential resources and data required to carry out this project.

Our sincere gratitude goes to the project review committee for their ongoing encouragement and guidance, which has enabled us to explore new areas of knowledge. We would also like to acknowledge the contributions of all faculty members of the department for their kind assistance and cooperation during the development of our project.

We would like to express our deep appreciation to our parents and family members, whose unwavering moral and financial support has been a constant source of strength throughout this journey.

Lastly, we wish to convey our sincere thanks to our friends, whose direct and indirect assistance has been invaluable in completing both the project and this report.

**Aayush Patil Ankit Doye Vaibhav Taywade (BT21CSE179) (BT21CSE170) (BT21CSE200)**

**Abstract**

Skin diseases, especially melanoma, represent a major global health issue due to their high prevalence and potential for severe outcomes if not detected early. Timely diagnosis of such diseases is crucial, as early intervention can significantly improve patient prognosis. Traditional methods of diagnosing skin lesions through visual inspection are challenging, as many lesions share similar characteristics. In recent years, deep learning techniques have shown promise in automating the detection of skin diseases, particularly through Convolutional Neural Networks (CNNs). However, CNNs are often limited in their ability to capture long-range spatial relationships in images, which are essential for distinguishing subtle variations in skin lesions. To address this limitation, the Vision Transformer (ViT) has emerged as a powerful alternative, utilizing a self-attention mechanism to better understand global image dependencies and improve classification accuracy.

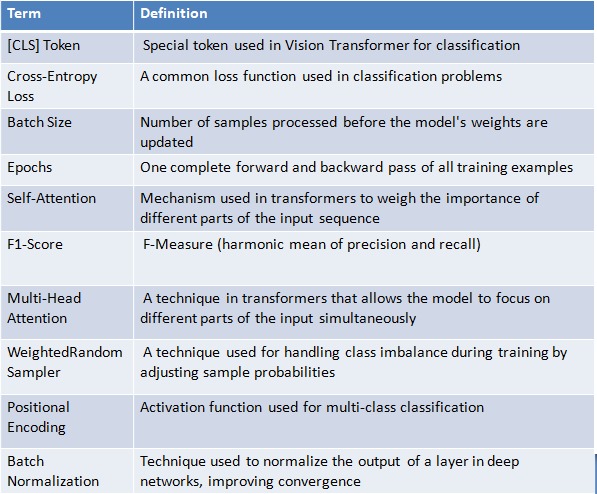
This study explores the application of ViT for skin disease detection using a large dataset of 35,000 images, categorized into 14 distinct classes. Among these, melanoma detection is a critical focus due to its role in early cancer diagnosis. Initially, we experimented with a CNN-based ResNet model; however, it failed to deliver satisfactory accuracy for this task. Consequently, we implemented the Vision Transformer, which showed significant improvements in model performance, particularly in classifying melanoma and other skin lesions. The ViT-based model was able to effectively identify relevant features across the images, overcoming the limitations of CNNs in capturing long-range spatial dependencies.

The results demonstrate that the Vision Transformer outperforms traditional CNN models, offering a more accurate solution for skin disease classification. By automating the process of skin lesion detection, this approach holds the potential to assist dermatologists in making faster and more accurate diagnoses, ultimately leading to better patient outcomes. This study highlights the advantages of transformer-based architectures for medical image analysis, particularly in tasks involving complex image patterns, and paves the way for future advancements in automated skin disease detection.

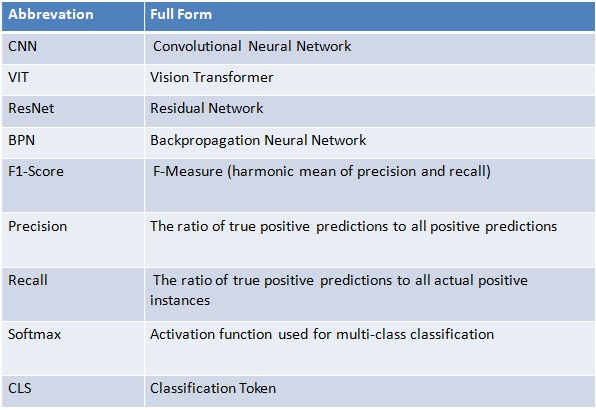
**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SR. NO** | **TOPICS** | **PAGE** |
| **1** | **Introduction**  1.1 Motivation For The Work  1.2 The Objective Of The Work  1.3 Scope And Significance Of Work  1.4 Organization Of The Thesis | **9** |
| **2** | **Literature Review** | **16** |
| **3** | **Work Done**  3.1 Methodology  3.2 Implementation of CNN Model in Image Classification  3.3 Implementation of Vision Transformer  3.4 Workflow | **21** |
| **4** | **Results And Discussion** | **36** |
| **5** | **Conclusion** | **42** |
| **6** | **References** | **46** |

**Definition Of Technical Terms**



**List Of Abbreviations**



**Chapter 1**

**Introduction**

Skin diseases, especially melanoma, are a growing health concern worldwide. Early detection is vital for effective treatment and improving survival rates, as conditions like melanoma can become life-threatening if diagnosed at later stages. Traditional methods of diagnosing skin lesions often rely on visual inspection by trained dermatologists, which can be prone to human error due to the subtle differences between various types of lesions. As the global incidence of skin diseases continues to rise, there is a pressing need for automated systems that can assist healthcare professionals in providing timely and accurate diagnoses.

In recent years, deep learning models have emerged as powerful tools for automating medical image analysis. Convolutional Neural Networks (CNNs), widely used in image classification tasks, have been applied to skin disease detection with some success. However, CNNs have limitations when it comes to capturing long-range spatial dependencies, which are crucial for identifying complex features in medical images. To overcome these limitations, Vision Transformers (ViT), a more recent innovation in deep learning, have been proposed. ViT utilizes self-attention mechanisms to better capture global relationships within images, making it a promising alternative for skin lesion classification.

The goal of this project is to develop an automated skin disease detection system using deep learning, with a focus on melanoma detection. We leverage a dataset containing 35,000 images categorized into 14 classes, including melanoma, to train and evaluate the performance of a Vision Transformer model. By implementing ViT, we aim to improve the accuracy of skin lesion classification compared to traditional CNN-based models, with the ultimate goal of creating a tool that can assist dermatologists in early detection, leading to better patient outcomes.

This report discusses the methodology, results, and challenges encountered during the development of the skin disease detection system, highlighting the potential of Vision Transformers in medical image analysis.

**1.1 Motivation for work**

Skin diseases, especially melanoma, have become a significant global health concern due to their high incidence rates and the potential for serious health consequences if not diagnosed and treated early. Melanoma is one of the most aggressive forms of skin cancer, and its early detection is critical for improving survival rates. Despite the advancements in medical imaging and diagnostic techniques, the process of diagnosing skin lesions remains challenging. Traditional methods, which rely on visual inspection by dermatologists, are subjective and prone to human error, particularly when distinguishing between benign and malignant lesions. This can result in delayed diagnoses and suboptimal treatment outcomes.

With the growing burden of skin diseases and the need for accurate and timely diagnosis, there is an increasing demand for automated systems that can assist healthcare professionals in identifying skin lesions. Machine learning and deep learning, especially Convolutional Neural Networks (CNNs), have shown great promise in automating medical image analysis tasks, including skin disease detection. However, CNNs often struggle with capturing long-range dependencies within images, which are crucial for distinguishing subtle differences in skin lesions. This limitation motivates the exploration of more advanced architectures like the Vision Transformer (ViT), which utilizes self-attention mechanisms to better understand global image features and long-range spatial relationships.

The motivation behind this work is twofold: to improve the accuracy and efficiency of skin disease detection through deep learning models, and to contribute to the development of automated tools that can assist dermatologists in providing faster and more accurate diagnoses. By leveraging the Vision Transformer, this project aims to address the limitations of traditional CNNs and demonstrate the potential of transformer-based models for complex medical image analysis tasks. Ultimately, this research aspires to make a meaningful contribution to the early detection of melanoma and other skin diseases, improving healthcare outcomes and reducing the burden on medical professionals.

**1.2 The Objective for Work**

The primary objective of this project is to develop an automated system for the detection and classification of skin diseases, with a particular focus on melanoma, using deep learning techniques. The specific objectives of the work are as follows:

1. **To Implement a Vision Transformer (ViT) Model**:  
   To explore and implement the Vision Transformer architecture for skin disease detection, leveraging its ability to capture long-range spatial dependencies in images, which is crucial for accurate lesion classification.
2. **To Improve Classification Accuracy**:  
   To evaluate the performance of the ViT model in classifying skin diseases and compare its accuracy to traditional Convolutional Neural Networks (CNNs), particularly in identifying melanoma and other skin lesions.
3. **To Utilize a Large Skin Disease Dataset**:  
   To train and evaluate the model on a dataset consisting of 35,000 images categorized into 14 distinct classes, ensuring the model can effectively generalize across various types of skin lesions.
4. **To Develop a Tool for Early Cancer Detection**:  
   To create an automated tool that can assist healthcare professionals in the early detection of melanoma, ultimately aiding in early cancer diagnosis and improving patient outcomes.
5. **To Contribute to Advancements in Medical Image Analysis**:  
   To contribute to the growing body of work on deep learning applications in medical image analysis, demonstrating the effectiveness of Vision Transformers in a domain where accuracy and early detection are critical.

By achieving these objectives, this project aims to develop a robust, accurate, and efficient model for skin disease detection, with a focus on melanoma, and to provide a scalable solution for real-world medical applications.

**1.3 Scope and Significance of Work**

**1.3.1 Scope**

The primary focus of this project is to develop a reliable automated system for identifying skin diseases, with special emphasis on melanoma, a critical form of skin cancer. The work involves the following key components:

1. **Dataset and Classification**:  
   A dataset comprising 35,000 images representing 14 different skin disease categories, including melanoma, is used to train and test the model. The classification task revolves around accurately categorizing these images.
2. **Model Selection**:  
   This study primarily investigates the Vision Transformer (ViT) model. Its performance is compared to a Convolutional Neural Network (ResNet), providing insights into which approach better handles this type of data.
3. **Evaluation Metrics**:  
   Several metrics, such as accuracy, recall, and precision, are employed to measure the effectiveness of the models. These evaluations help ensure that the system can distinguish between different skin diseases effectively.
4. **Data Preparation**:  
   Images in the dataset are preprocessed through resizing, normalization, and augmentation techniques. These steps are necessary to handle variations in the data and enhance the model’s ability to generalize.
5. **Technology Stack**:  
   The implementation uses frameworks like PyTorch and TensorFlow, with additional tools for metrics and data visualization.

This project is primarily research-oriented and focuses on evaluating the capabilities of Vision Transformers for medical image analysis tasks.

**1.3.2 Significance**

This work is significant due to its implications in healthcare and medical imaging. Below are some of the major contributions and impacts of the project:

1. **Early Detection for Better Outcomes**:  
   Melanoma is aggressive but treatable if caught early. An automated tool that quickly identifies potential cases can lead to timely interventions and improved survival rates.
2. **Support for Healthcare Workers**:  
   Dermatologists often face challenges in handling a large number of cases. This project provides a potential solution to assist medical professionals by offering preliminary insights, reducing their workload, and enabling faster diagnoses.
3. **Better Accuracy with New Technology**:  
   Vision Transformers introduce a fresh perspective to image analysis by understanding long-range relationships in images. This capability improves classification accuracy compared to traditional methods like CNNs.
4. **Application Potential**:  
   While this project is primarily academic, it lays the foundation for developing tools that can eventually be used in clinical practice, particularly in regions with limited access to trained specialists.
5. **Contribution to Research**:  
   By applying an emerging model like ViT to a practical medical problem, this work pushes the boundaries of existing research and inspires further exploration in similar fields.
6. **Scalability and Relevance**:  
   The techniques used here are scalable and can be adapted to other medical imaging tasks, opening possibilities for a wider range of healthcare applications.

This project not only addresses a specific need in medical imaging but also demonstrates the potential of advanced machine learning models in solving real-world challenges. It emphasizes both academic contributions and practical implications, making it a meaningful step in the evolution of healthcare technology.

**1.4 Organization of the Thesis**

This thesis is organized into the following chapters:

* **Chapter 1: Introduction**The first chapter provides an overview of the project, highlighting the importance of skin disease detection, especially melanoma, and the challenges involved in traditional methods of diagnosis. It also introduces the use of deep learning models, specifically the Vision Transformer, to address these challenges. The chapter outlines the motivation behind the project, the dataset used, and the goals of the research.
* **Chapter 2: Literature Review**Chapter 2 presents a comprehensive review of the existing literature related to skin disease detection, including both traditional methods and deep learning approaches. This chapter discusses the role of Convolutional Neural Networks (CNNs) in medical image analysis, the emergence of Vision Transformers, and their application in various domains, particularly in medical image segmentation and classification. Relevant studies, their findings, and the gaps identified are also discussed in this chapter.
* **Chapter 3: Methodology**Chapter 3 details the methodology employed in this project, including the data collection process, preprocessing steps, and the model architecture. It focuses on the Vision Transformer (ViT), explaining how it is implemented for skin disease detection, and compares its performance to other models such as ResNet. The chapter also outlines the training process, evaluation metrics, and the software tools and frameworks used.
* **Chapter 4: Results and Discussion**In Chapter 4, the results of the experiments are presented and analyzed. This chapter discusses the performance of the Vision Transformer model in comparison to the initial CNN model (ResNet) in terms of accuracy and other relevant metrics. The discussion includes insights into the model's strengths and limitations, as well as potential areas for improvement.
* **Chapter 5: Conclusion and Future Work**Chapter 5 summarizes the findings of the project and provides conclusions drawn from the results. This chapter also suggests potential avenues for future research, improvements in model performance, and applications of the developed system in real-world healthcare settings.
* **Chapter 6: References**Chapter 6 provides a comprehensive list of all the references cited throughout the thesis. It includes scholarly articles, books, conference papers, and other resources that contributed to the research and development of this project.

Each chapter builds on the previous one, providing a comprehensive understanding of the project, from motivation and background to methodology, results, and potential future directions.

**Chapter 2**

**LITERATURE REVIEW**

**[1] SCI\_ Machine learning approach for classification of maculopapular and vesicular rashes using the textural features of the skin images**

The research paper presents a machine learning-based approach to classify maculopapular and vesicular rashes using textural features extracted from skin images. A backpropagation neural network (BPN) model is employed, focusing on distinguishing these two types of rashes through features derived from the Gray-Level Co-Occurrence Matrix (GLCM). Key features include contrast, correlation, energy, and homogeneity, which characterize the texture of the rash regions.

The methodology involves two phases: using unsegmented images and applying segmentation to isolate the rash regions. Segmentation techniques such as the Otsu method and K-Means clustering enhance classification accuracy by refining the focus on the rash areas. Images were preprocessed, resized, and used to train and test the neural network. The study evaluates the model's performance using metrics such as accuracy, sensitivity, specificity, and F1-score.

The results indicate an average accuracy of 83.43% for segmented images, with improved sensitivity (92.39%) and F1-score (86%). This approach highlights the potential for integrating such models into teledermatology applications for remote healthcare.

**[2] Deep Learning in Skin Disease Image Recognition**

The paper explores the application of deep learning techniques for skin disease image recognition. It systematically reviews existing research from 2016 to 2020, analyzing datasets, preprocessing techniques, and model architectures commonly used in the field. Models such as ResNet, DenseNet, AlexNet, and VGG are prominently applied, with a focus on convolutional neural networks (CNNs) for feature extraction and classification. Multimodel fusion methods are highlighted as superior to single-model approaches, offering enhanced performance through the integration of diverse features.

The methodology emphasizes the use of data augmentation, preprocessing, and dataset fusion to address challenges like dataset scarcity and variability in skin disease images. Evaluation metrics such as accuracy, sensitivity, specificity, and AUC are employed to measure performance. Some studies report accuracy rates exceeding 90%, demonstrating the efficacy of deep learning in achieving dermatologist-level diagnosis.

This comprehensive review suggests future research directions, including the generation of synthetic datasets, improving model interpretability, and adapting algorithms for clinical environments, paving the way for AI-driven advancements in dermatology.

**[3] Deep Learning Based Dermatological Condition Detection A Systematic Review With Recent Methods Dataset**

This paper systematically reviews deep learning methods for detecting dermatological conditions through dermoscopic images, addressing diseases such as melanoma, acne, and eczema. It examines 22 deep learning models, including CNN-based architectures like VGG-16, MobileNet, DenseNet, ResNet, and GANs. The approaches typically involve preprocessing, segmentation, feature extraction, and classification. For example, ResNet-50 demonstrated high accuracy in melanoma classification, while MobileNet V2 combined with LSTM improved performance for imbalanced datasets.

Key datasets include ISIC, DermNet, and self-generated hospital datasets. Evaluation metrics used are accuracy, sensitivity, specificity, and others, achieving accuracies often exceeding 85% depending on the model and dataset.

Challenges highlighted include limited labeled data, dataset imbalances, lack of diversity in skin tones, and interpretability issues in deep learning. Future directions suggest leveraging more comprehensive datasets, integrating clinical metadata, and addressing rare dermatological conditions for broader applications and improved diagnostic reliability.

**[4] Skin Lesion Segmentation Based on Vision Transformers and**

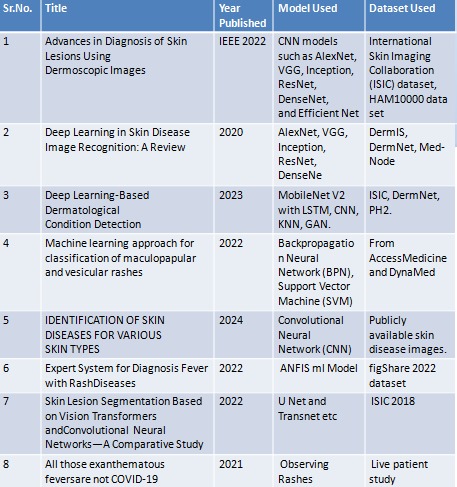
**Convolutional Neural Networks - A Comparative Study**

The paper systematically reviews advancements in deep learning for dermatological condition detection using dermoscopic images. It examines 22 methods, including CNNs, MobileNet, ResNet, DenseNet, and VGG models, used for classifying conditions like melanoma, psoriasis, and acne. The approaches include preprocessing (noise removal, augmentation), segmentation (region growing, edge detection), and classification (feature extraction and prediction). For example, ResNet-50 and Inception networks classify conditions with accuracy exceeding 85% on datasets like ISIC and DermNet. Transfer learning and ensemble models, like DenseNet-121 combined with MobileNetV1, enhance feature extraction and accuracy.

Key challenges include imbalanced datasets, limited diversity in skin types, and model interpretability issues. Future directions suggest incorporating metadata, improving data quality, and targeting underrepresented conditions such as burns and rare rashes.

This paper highlights the transformative potential of deep learning in dermatology while addressing critical gaps in inclusivity, interpretability, and performance across diverse clinical scenarios​(newset)

**Literature Review Summary:**



**Chapter 3**

**WORK DONE**

**3.1 Methodology**

This section outlines the methodology used to develop and train a **Vision Transformer (ViT)** model for the classification of skin diseases. The following steps were undertaken to design, implement, and evaluate the model.

#### **3.1.1 Dataset Collection and Preprocessing**

The dataset used for this project consists of skin disease images organized into 14 classes. The dataset is sourced from an image folder structure, with each folder corresponding to a specific class of skin disease. The dataset is divided into three sets:

* Training Set: Used for model training.
* Validation Set: Used for tuning model parameters and preventing overfitting.
* Test Set: Used for evaluating the model’s final performance.

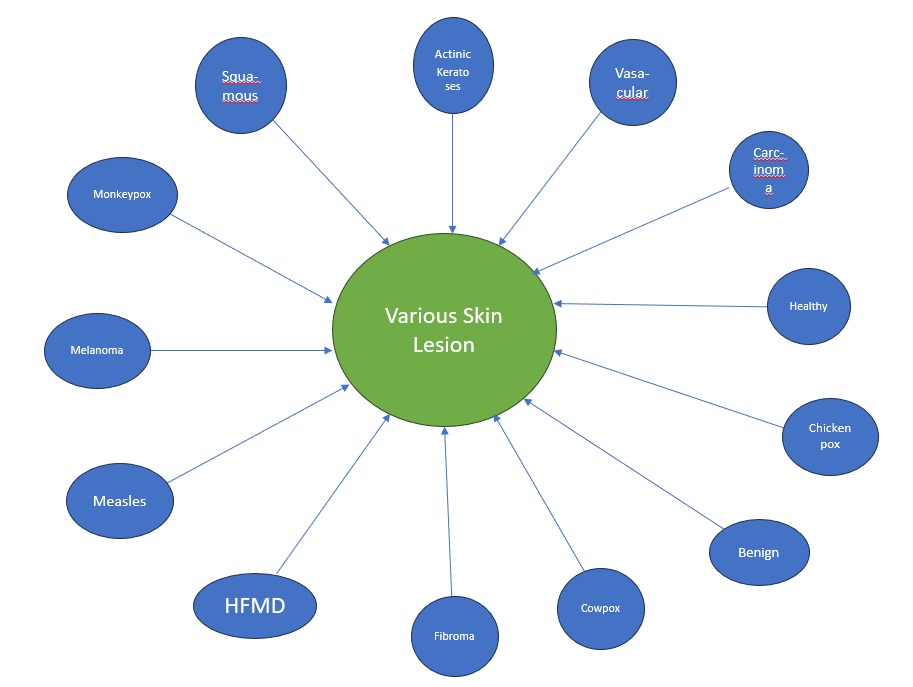
The following preprocessing steps were applied to the images:

* Resizing: All images were resized to a uniform size of 224x224 pixels to match the input size expected by the Vision Transformer (ViT) model.
* Normalization: Each image was normalized using the mean and standard deviation values from the pre-trained ViT model. This ensures that the input images are on the same scale as those used during the pre-training of the model.

#### **3.1.2 Data Augmentation and Class Balancing**

To enhance model generalization and address class imbalance, the following strategies were employed:

* Data Augmentation: Although not explicitly mentioned in the code, data augmentation can be used to introduce random transformations (e.g., rotations, flips) during training to make the model more robust to variations in skin lesion images.
* Class Balancing: Due to class imbalances in the dataset, a WeightedRandomSampler was used. This sampler adjusts the probability of selecting each class based on its frequency in the training dataset, ensuring that the model is exposed to all classes more evenly during training.



#### **3.1.3 Model Architecture**

The core of the model is the Vision Transformer (ViT), specifically the ViT-base-patch16-224-in21k model, pre-trained on the ImageNet dataset. The model is designed to handle image classification tasks by capturing long-range dependencies in images using a self-attention mechanism. The ViT model is fine-tuned on the skin disease dataset to classify images into one of 14 classes.

Key model parameters:

* Output Classes: 14 classes corresponding to different skin diseases.
* Transformer Configuration: Patch size of 16x16 and input size of 224x224 pixels.
* Loss Function: CrossEntropyLoss was used for multi-class classification.
* Optimizer: AdamW optimizer with a learning rate of 5e-5 was used for model optimization.

#### **3.1.4 Training Process**

The model was trained for 10 epochs using a batch size of 32. The training process is divided into the following steps:

* Forward Pass: During each epoch, the input images were passed through the ViT model, and the model output (predicted class logits) was computed.
* Loss Calculation: The CrossEntropyLoss function was used to calculate the loss between the predicted and true class labels.
* Backpropagation: Gradients were computed for the model’s parameters, and the optimizer adjusted the model weights to minimize the loss.
* Metrics Calculation: The model’s training accuracy and validation accuracy were tracked after each epoch to monitor progress and prevent overfitting.

#### 

#### **3.1.5 Evaluation and Metrics**

After training, the model's performance was evaluated on the test set. The following metrics were used to assess the model’s effectiveness:

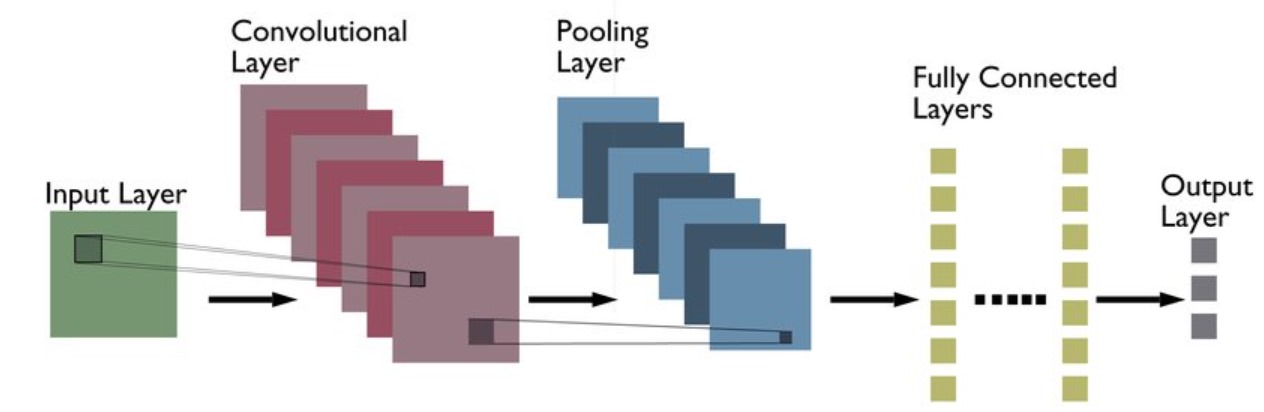
* Accuracy: The percentage of correctly classified images.
* Precision, Recall, and F1-score: These metrics were calculated using the classification\_report from the scikit-learn library. These metrics provide a detailed evaluation of the model’s performance for each class.
* Confusion Matrix: A confusion matrix was generated to visualize the distribution of predicted and actual classes. This helps in identifying where the model misclassified certain skin diseases.

#### **3.1.6 Results Visualization**

To visualize the model's training progress and performance:

* Loss Curve: The training and validation losses were plotted over the epochs to track the model’s learning curve.
* Accuracy Curve: The training and validation accuracies were plotted to observe how well the model generalizes on the validation data.
* Confusion Matrix: A heatmap of the confusion matrix was displayed to show the true positive and false positive rates for each class, highlighting which skin diseases were more challenging for the model to classify.

**3.2 Implementation of CNN Model in Image Classification (ResNet)**



#### Initially, we implemented a Convolutional Neural Network (CNN) using the ResNet architecture to classify skin diseases. ResNet, or Residual Networks, is a deep learning model known for utilizing residual connections, which help train very deep networks without suffering from vanishing gradients. This feature allows ResNet to learn more complex patterns in images, making it a widely-used architecture for image classification tasks, including skin disease detection.

#### **3.2.1 ResNet Architecture:**

#### Input Layer: The model accepts input images of size 224x224x3 (height x width x channels), which is the standard input size for models like ResNet.

#### Convolutional Layers: The initial layers of ResNet apply various filters to the image in order to detect edges, textures, and basic patterns.

#### Residual Blocks: The network uses residual blocks where skip connections are employed, allowing the input to bypass certain layers and be added to the output of the subsequent layers. These skip connections help alleviate the vanishing gradient problem, allowing deeper networks to be trained effectively.

#### Residual Connection Formula: The output of a residual block is given by: Output=Activation(Input+Residual)\text{Output} = \text{Activation}(\text{Input} + \text{Residual})Output=Activation(Input+Residual)

#### These residual connections allow the network to learn more complex features at deeper layers without losing information.

#### Pooling Layers: After convolutional operations, max-pooling layers are used to reduce the spatial dimensions (height and width) of the feature maps while preserving important features. This also reduces the computational cost.

#### Fully Connected (Dense) Layers: The feature maps from the convolutional and pooling layers are flattened into a 1D vector and passed through fully connected layers. These layers are responsible for making the final classification decision based on the extracted features.

#### Output Layer: The last layer of the network produces logits (raw scores) for each class, and a softmax function is applied to generate class probabilities. The class with the highest probability is chosen as the predicted label.

#### **3.2.2 Training the ResNet Model:**

#### Training Duration: The ResNet model was trained for 50 epochs on the skin disease dataset.

#### Loss Function: Cross-Entropy Loss was used as the loss function, as it is well-suited for multi-class classification tasks.

#### Optimizer: The Adam optimizer was used for training, which is known for adapting the learning rate for each parameter and is generally efficient for deep learning models.

#### Training Process:

#### During each epoch, the model was trained using mini-batches of images.

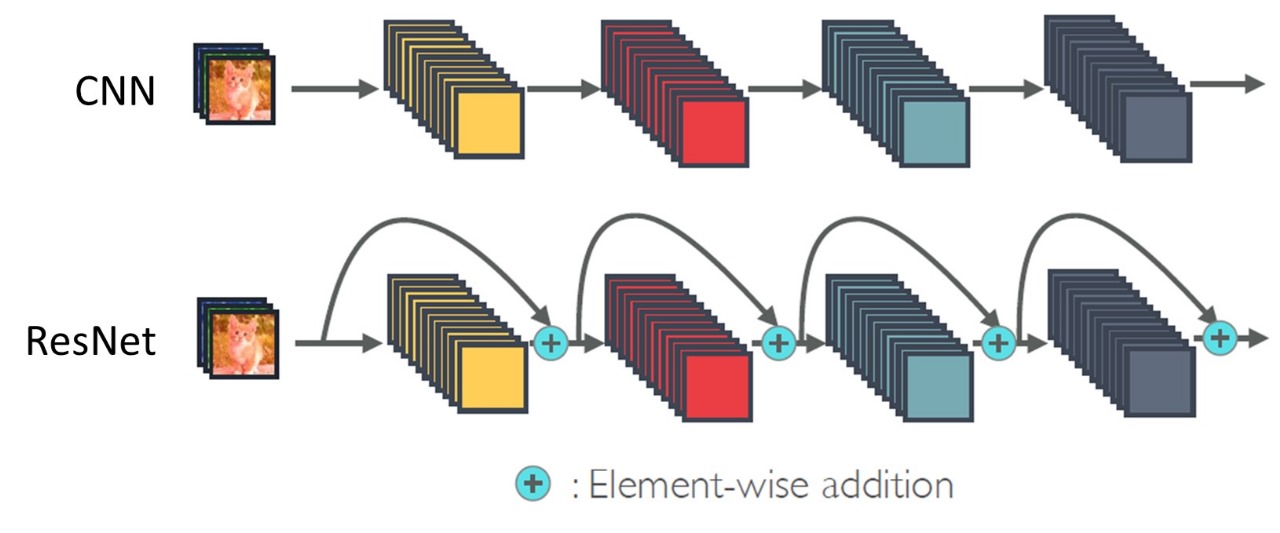
#### For each batch, the model performed a forward pass to calculate predictions, and the loss was computed by comparing the predicted labels with the true labels.

#### Backpropagation was used to adjust the weights of the network, aiming to minimize the loss function.

#### The accuracy of the model was recorded at the end of each epoch to track its progress.

#### **3.2.3 Performance and Accuracy of the ResNet Model**

#### Despite training the ResNet model for 50 epochs, the performance was unsatisfactory with a maximum accuracy of 77.7% on the validation set. The model was unable to achieve higher accuracy due to its inherent limitations in capturing global dependencies between image features. Skin disease classification, especially for complex diseases like melanoma, requires a model that can capture long-range relationships within the image. ResNet’s reliance on local receptive fields limited its ability to distinguish subtle differences between similar skin lesions, which led to suboptimal performance.

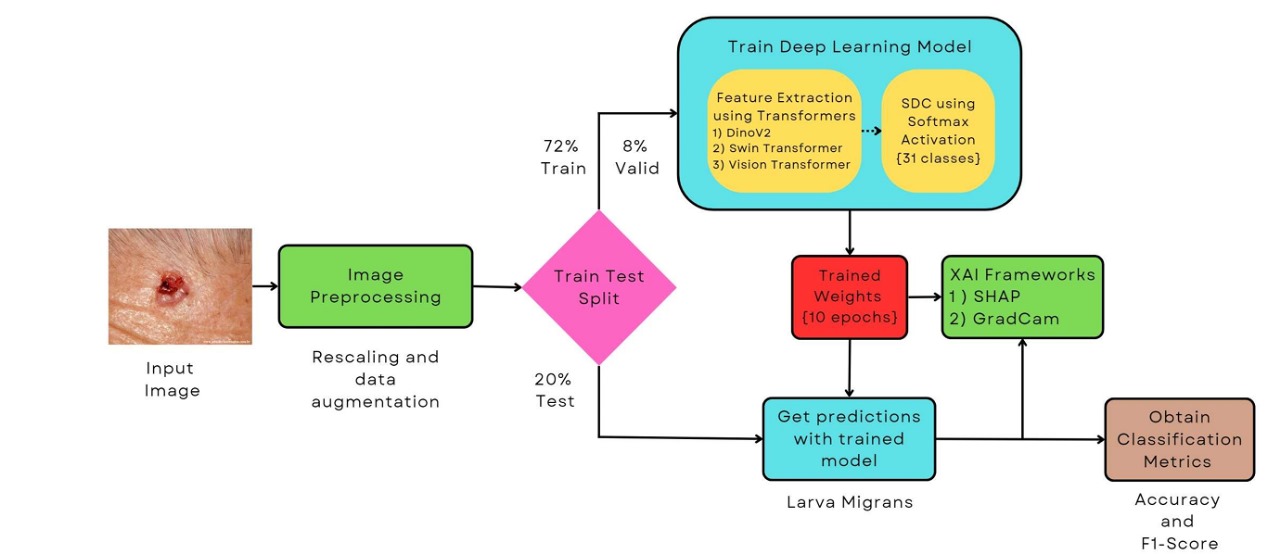


#### **3.2.4 Conclusion**

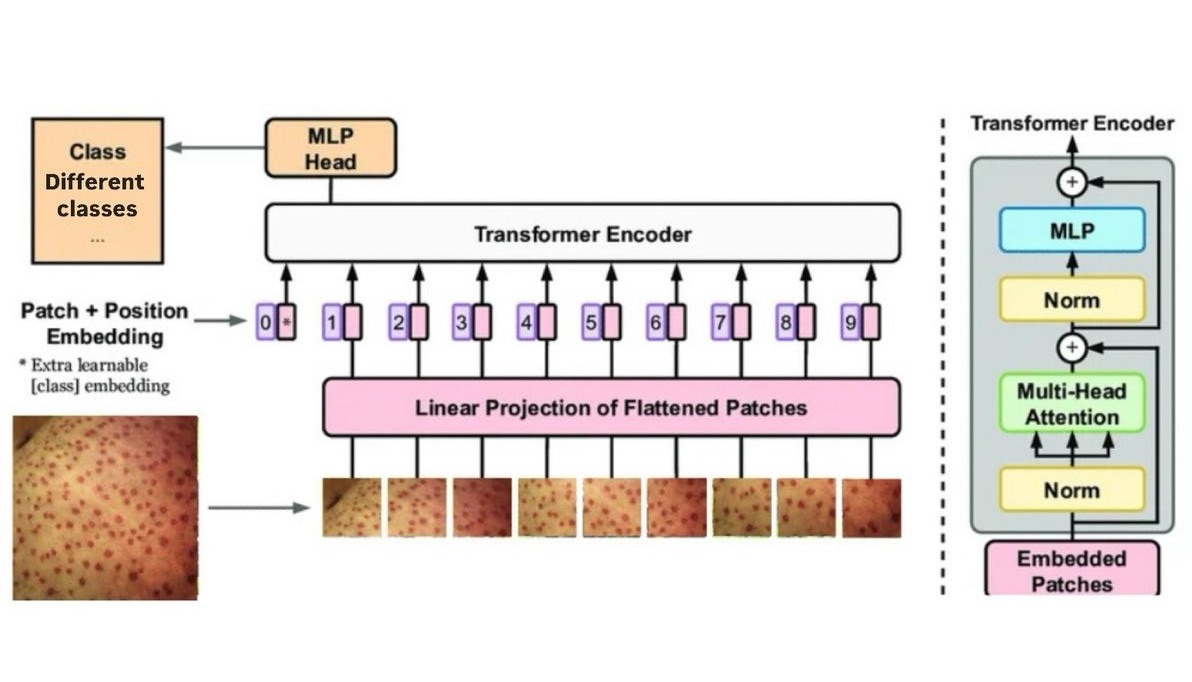
#### Although the ResNet-based CNN model was a solid starting point, its performance was insufficient with 77.7% accuracy after training for 50 epochs. The low accuracy was primarily due to the model’s limitations in capturing long-range spatial dependencies. The switch to Vision Transformer (ViT), which utilizes a more advanced approach of processing image patches and learning global relationships, resulted in significant performance improvements.

**3.3 Implementation of Vision Transformer**

After the ResNet model demonstrated unsatisfactory performance with an accuracy of 77.7% after 50 epochs, we transitioned to using the Vision Transformer (ViT) model for skin disease classification. ViT is based on the transformer architecture, which was originally designed for natural language processing (NLP) tasks but has been successfully adapted for image classification. ViT has shown promising results in computer vision tasks due to its ability to capture long-range dependencies and learn global relationships between different parts of the image.

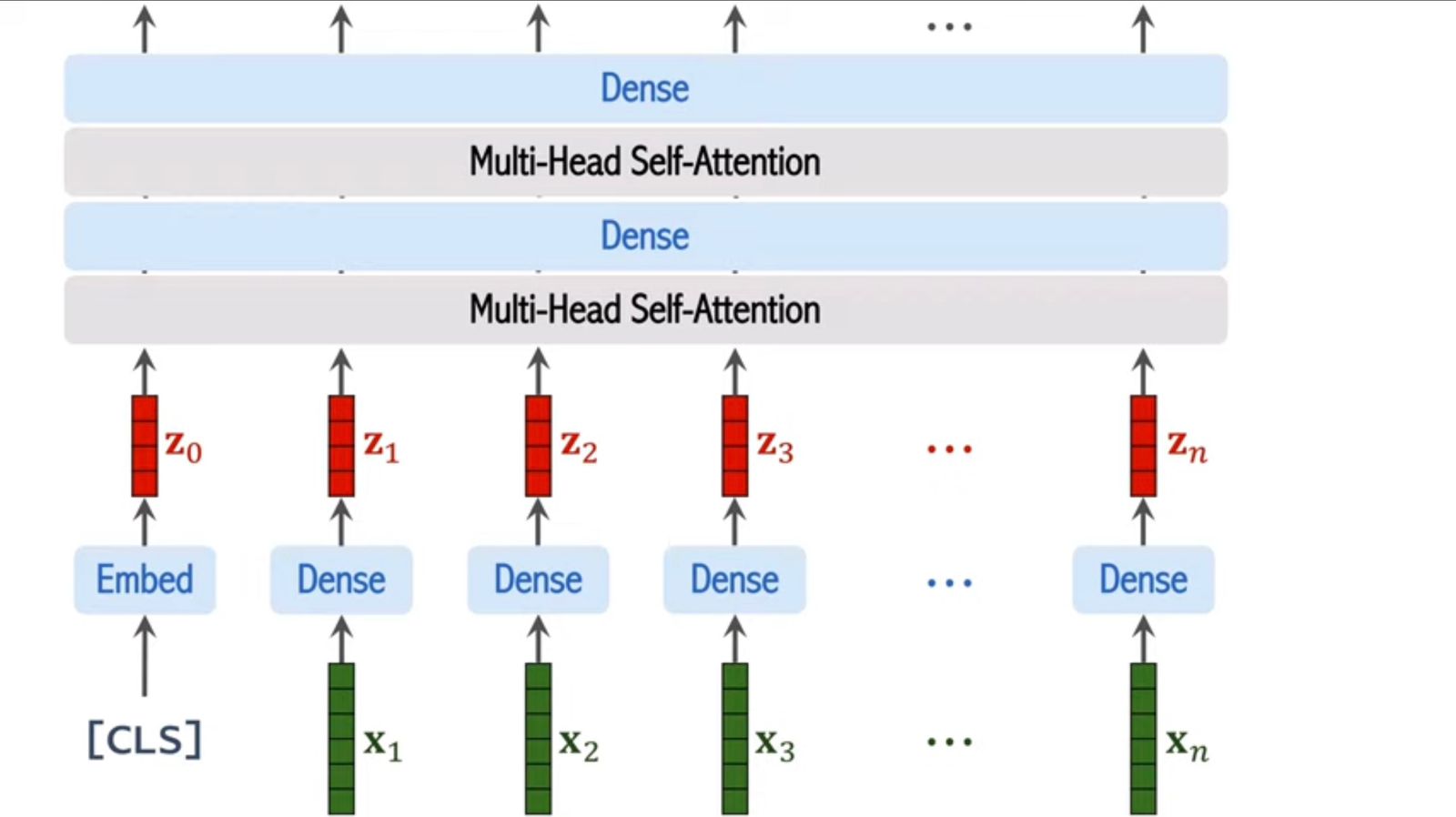


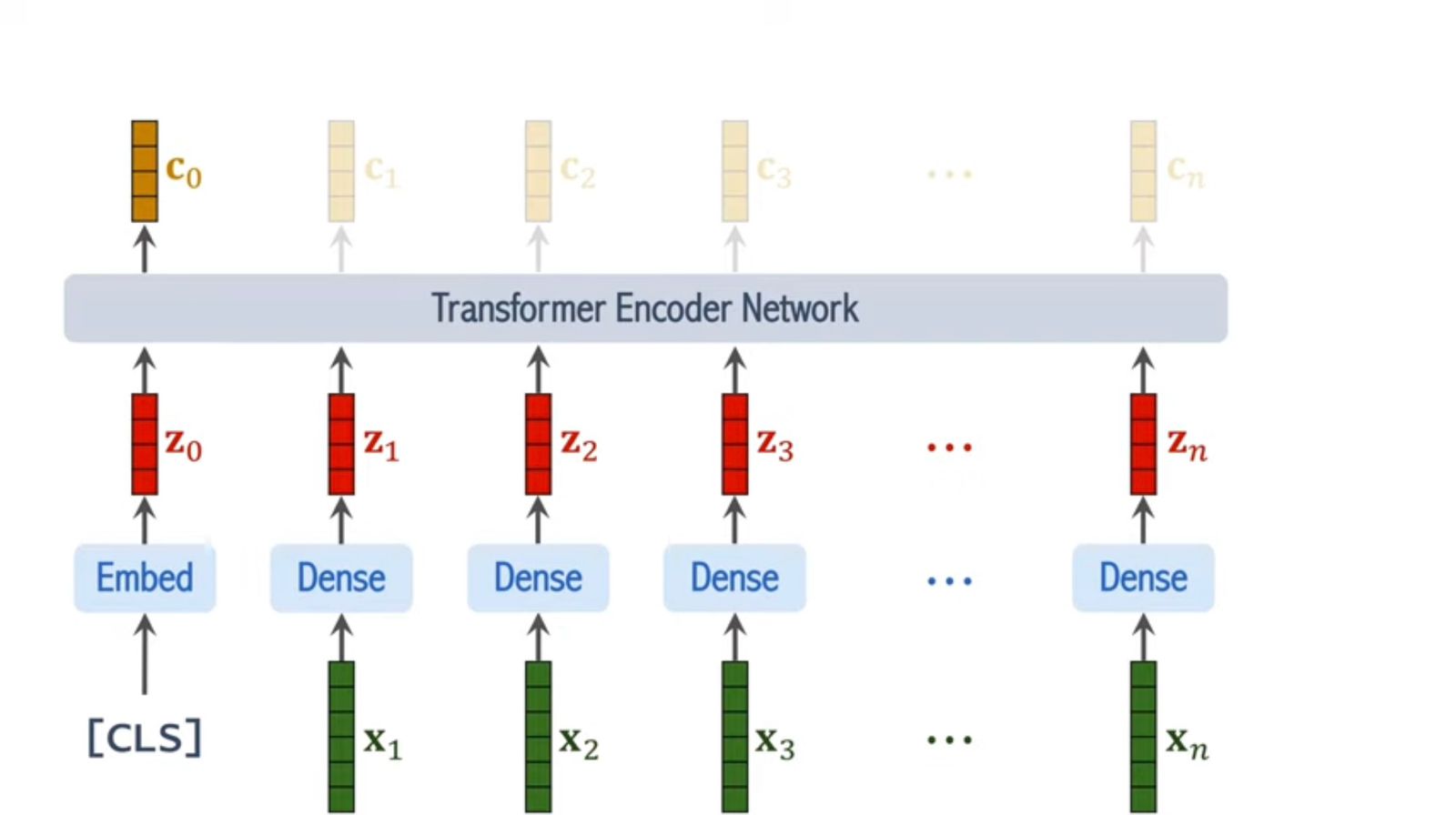
**3.3.1 ViT Architecture and Functioning:**

* Input Representation: Unlike traditional Convolutional Neural Networks (CNNs), ViT first splits the input image into small, fixed-size patches. These patches are flattened into one-dimensional vectors and linearly embedded into a higher-dimensional space, forming a sequence of patch embeddings. For example, an image of size 224x224 pixels is divided into patches of size 16x16 pixels, resulting in 196 patches (since 22416=14\frac{224}{16} = 1416224​=14).
* Positional Encoding: Since transformers do not have an inherent sense of spatial relationships, positional encoding is added to each patch embedding to provide information about the position of the patch in the original image.
* Transformer Encoder: The patches, along with positional encodings, are passed through a transformer encoder consisting of multiple layers of multi-head self-attention and feedforward networks. The attention mechanism allows each patch to interact with all other patches in the image, which helps the model capture long-range dependencies. Each transformer block processes the sequence of patch embeddings, and through self-attention, learns relationships between patches across the image.
  + Multi-head Attention: This allows the model to focus on different parts of the image simultaneously, learning different types of relationships (e.g., spatial, texture, context).
  + Feedforward Networks: These networks help transform the output from the attention mechanism into a higher-level representation.
* Classification Head: After passing through the transformer encoder layers, the final output is a vector corresponding to the [CLS] token. This token aggregates information from all the patches and is used for classification. The output of the [CLS] token is passed through a fully connected (FC) layer to produce logits for each class.
  + Softmax Activation: A softmax function is applied to the logits to generate class probabilities. The class with the highest probability is selected as the predicted label.

#### **3.3.2 Training the ViT Model**

* Training Configuration: The ViT model was trained for 10 epochs with a batch size of 32. The model was initialized with pre-trained weights from the ViT-base-patch16-224-in21k model, which had been trained on a large-scale dataset (ImageNet). The model was fine-tuned on our skin disease dataset to classify 14 different skin disease classes.
* Loss Function: We used Cross-Entropy Loss, which is appropriate for multi-class classification tasks, where the goal is to minimize the difference between predicted class probabilities and the true labels.
* Optimizer: The AdamW optimizer with a learning rate of 5e-5 was used to update the model's parameters. AdamW is known for its efficiency in fine-tuning large pre-trained models like ViT.

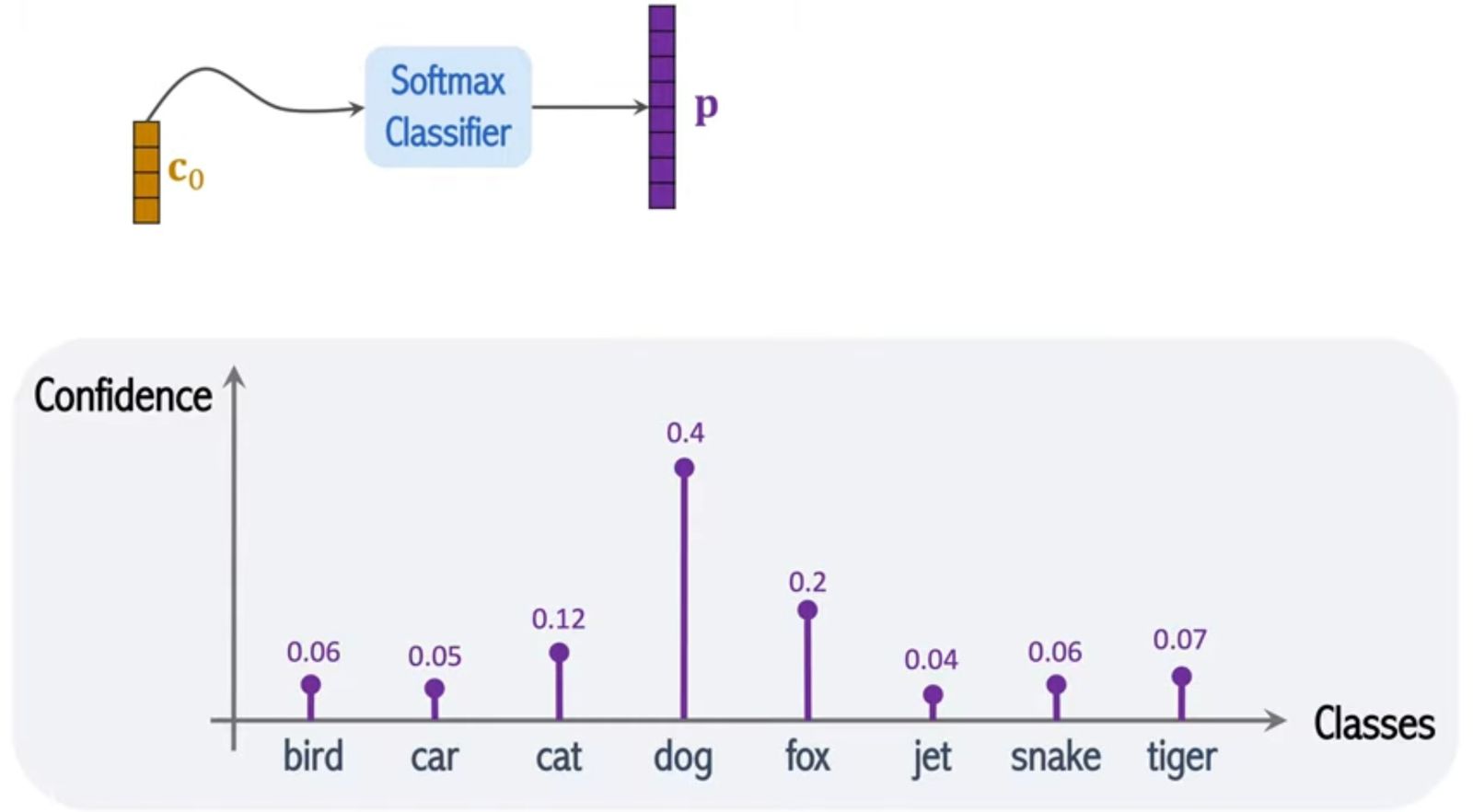




#### 

#### 

#### Thus, the softmax classifier will produce a vector containing 14 cells in our case, each representing a confidence score. The cell with the highest confidence score will correspond to the predicted class.

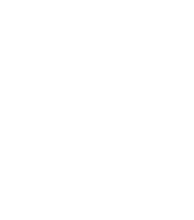


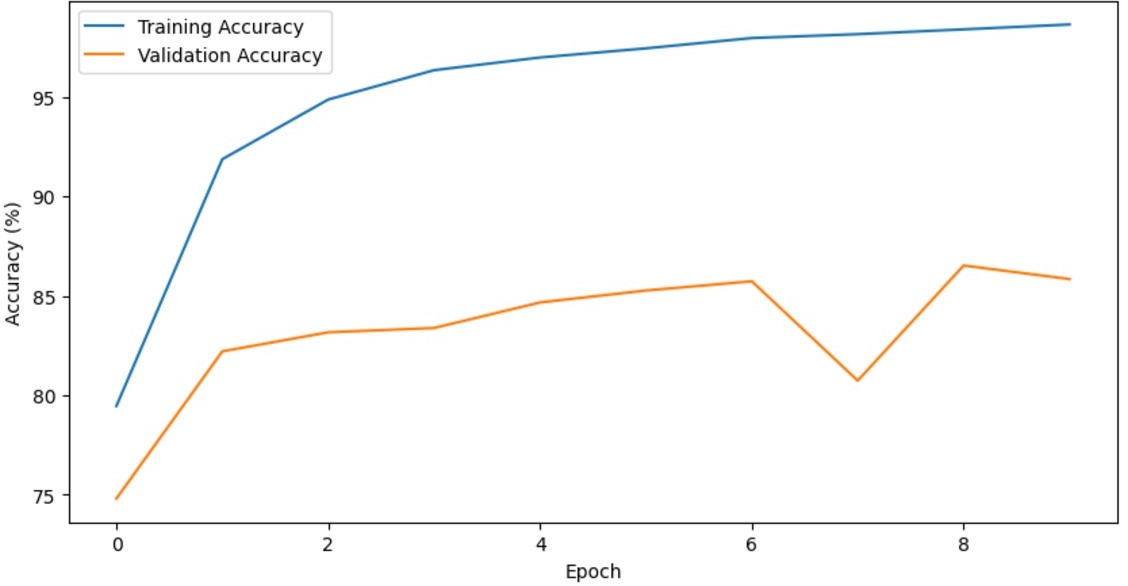
#### **3.3.3 Evaluation of the ViT Model**

After training the ViT model for 10 epochs, we observed a significant improvement in the model's performance compared to the ResNet model. The training accuracy reached 98.66%, indicating that the model was learning the features of the skin diseases well during training. The test accuracy achieved was 86.53%, which is a substantial improvement over the ResNet model’s 77.7%.

* Training Accuracy: 98.66% (after 10 epochs)
* Testing Accuracy: 86.53% (on the test set)

This performance suggests that the ViT model was better able to capture the global relationships between patches in the images, leading to improved classification of skin diseases. The higher testing accuracy also indicates that the model generalizes well to unseen data, which is critical for real-world applications.







#### **3.3.4 Conclusion**

The transition from the ResNet-based CNN model to the Vision Transformer (ViT) model significantly improved the classification accuracy of the skin disease dataset. The ViT model achieved a training accuracy of 98.66% and a testing accuracy of 86.53%, demonstrating its effectiveness in handling complex image classification tasks. The ability of ViT to capture long-range dependencies and global context within the image helped it achieve superior performance compared to traditional CNN models.

**3.4 Workflow**

The workflow for the skin disease classification project using Vision Transformer (ViT) involved several systematic steps, from dataset preparation and model implementation to training and evaluation. The following outlines the work process:

#### **3.4.1 Dataset Collection and Organization**

* Dataset: The dataset used for the project consists of skin disease images categorized into 14 classes, each representing a different type of skin condition. The dataset is organized in a folder structure where each folder corresponds to one class.
* Data Split: The dataset was divided into three parts:
  + Training Set: Used to train the model.
  + Validation Set: Used for model tuning and hyperparameter selection.
  + Test Set: Used for final evaluation of the model's performance.

#### **3.4.2 Data Preprocessing**

* Resizing: The images were resized to a standard dimension of 224x224 pixels to match the input size required by the models (ResNet and ViT).
* Normalization: Images were normalized using the mean and standard deviation of the pre-trained ViT model, ensuring the images had the same scale as those used for training the pre-trained model.
* Class Balancing: Due to the class imbalance, a WeightedRandomSampler was applied to adjust the sampling frequency of each class, ensuring that the model was exposed to each class evenly during training.

#### **3.4.3 Implementation of ResNet (CNN) Model**

* Model Architecture: Initially, a ResNet-based CNN model was implemented for the classification task. ResNet is a deep convolutional network with residual connections designed to train deep models effectively without losing important information during backpropagation.
* Training: The ResNet model was trained for 50 epochs using Cross-Entropy Loss and Adam optimizer. The training performance was monitored by tracking accuracy and loss during each epoch.
* Evaluation: After training, the model achieved a 77.7% accuracy on the test set. While the model showed reasonable performance, the accuracy was considered too low for the reliable detection of skin diseases.

#### **3.4.4 Transition to Vision Transformer (ViT)**

* Motivation for Change: Given the limitations of the ResNet model, the decision was made to switch to the Vision Transformer (ViT) model. ViT was chosen for its ability to capture global relationships between image patches, making it more effective for complex image classification tasks like skin disease detection.
* Model Architecture: The ViT-base-patch16-224-in21k model was used, which is a pre-trained version of ViT. The image was divided into non-overlapping patches, each of size 16x16 pixels, which were then flattened and passed through the transformer layers.
* Fine-Tuning: The pre-trained ViT model was fine-tuned on the skin disease dataset for 10 epochs. The model was trained using the AdamW optimizer with a learning rate of 5e-5 and Cross-Entropy Loss as the loss function.

#### **3.4.5 Training and Validation of ViT**

* Training Loop: The ViT model was trained for 10 epochs with a batch size of 32. During each epoch:
  + The model was fed batches of images, and the forward pass was performed to calculate predictions.
  + The backpropagation algorithm was used to adjust the model's weights by minimizing the loss.
* Validation: After each epoch, the model’s performance was evaluated on the validation set. The validation accuracy was recorded, and adjustments were made to ensure the model wasn't overfitting to the training data.

#### **3.4.6 Evaluation of ViT Model**

* Final Evaluation: After 10 epochs of training, the ViT model achieved a training accuracy of 98.66% and a test accuracy of 86.53%.
* Metrics: The performance was evaluated using key metrics such as accuracy, precision, recall, and F1-score for each class (skin disease).
* Confusion Matrix: A confusion matrix was generated to visualize the distribution of correct and incorrect classifications across the 14 skin disease classes. This helped identify which classes were being misclassified and provided insights into potential areas for improvement.

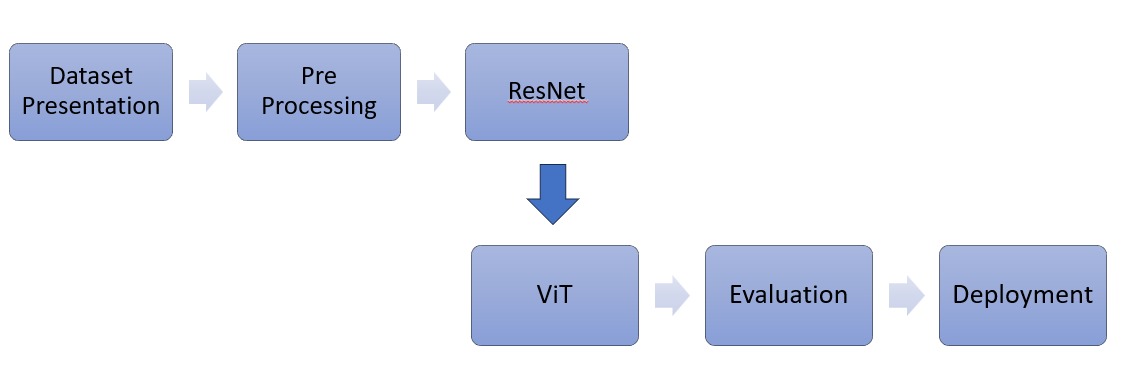
#### 

#### **3.4.7 Results Visualization**

* Loss and Accuracy Curves: To visualize the model’s performance, plots were generated showing the training and validation loss and training and validation accuracy over the 10 epochs. These plots indicated that the model was converging well and not overfitting.
* Confusion Matrix Visualization: A heatmap of the confusion matrix was plotted to show the model’s performance for each skin disease class, with annotations indicating the number of true positives and false positives.

#### **3.4.8 Model Deployment and Future Work**

* Once trained, the ViT model demonstrated promising accuracy and could potentially be integrated into an automated skin disease detection system. Future work may involve:
  + Model Fine-tuning: Further fine-tuning with additional data or applying techniques like data augmentation to improve generalization.
  + Real-Time Deployment: Developing a mobile or web-based application for real-time skin disease classification, especially in telemedicine and remote healthcare settings.



**Chapter 4**

**Results And Discussion**

In this chapter, we present and analyze the results of the skin disease classification task using both Convolutional Neural Networks (ResNet) and Vision Transformers (ViT). The performance of the models is discussed in terms of accuracy, precision, recall, and F1-score, with a particular focus on the improvements achieved by transitioning from ResNet to ViT.

#### **4.1 Overview of Model Performance**

The primary goal of this project was to develop a deep learning model capable of classifying skin diseases from a dataset containing 14 classes. Initially, we employed the ResNet architecture, a popular Convolutional Neural Network (CNN) model, for this task. After training the ResNet model for 50 epochs, we found that the model's accuracy was limited, reaching only 77.7% on the test set. Given the challenges faced with this architecture, we transitioned to the Vision Transformer (ViT) model, which achieved significantly improved results.

The ViT model was fine-tuned for 10 epochs and achieved a training accuracy of 98.66% and a test accuracy of 86.53%, reflecting the model's ability to capture complex global relationships within the images and improve classification performance.

#### **4.2 Performance Evaluation of ResNet (CNN)**

The ResNet model was trained on the skin disease dataset for 50 epochs. Despite the effectiveness of Residual Networks (ResNet) in general, the performance on this particular task was suboptimal, with the test accuracy plateauing at 77.7%. Below is a breakdown of the performance:

* Training Accuracy: The ResNet model showed a good fit on the training data, but due to its inability to capture global context, it struggled with accurately classifying more complex skin lesions, especially those that were subtle in appearance.
* Validation Accuracy: The model's accuracy on the validation set was also lower, indicating that it was overfitting to the training data and not generalizing well to unseen images.
* Loss: The loss continued to decrease over time, but the accuracy improvements were marginal, further suggesting that the model was struggling to differentiate between certain classes.

**4.3 Challenges with ResNet:**

* Local Feature Focus: As a CNN, ResNet excels at capturing local features, but it has limitations in understanding the global context of an image. This is particularly critical in skin disease classification, where lesions may have complex spatial relationships that a CNN struggles to capture.
* Overfitting: Despite being trained for 50 epochs, the model overfitted the training data and did not generalize well to unseen data, leading to suboptimal performance.
* 

#### **4.4 Performance Evaluation of Vision Transformer (ViT)**

In response to the limitations of the ResNet model, we transitioned to the Vision Transformer (ViT), which showed significant improvements in performance. The ViT model, pre-trained on a large dataset (ImageNet), was fine-tuned on our skin disease dataset for 10 epochs. The results were as follows:

* Training Accuracy: The training accuracy of the ViT model reached an impressive 98.66%, indicating that the model was effectively learning to classify the skin disease images.
* Test Accuracy: The test accuracy achieved by the ViT model was 86.53%, a substantial improvement over the ResNet model's performance. This indicates that the ViT model was better at generalizing to new, unseen data and making accurate predictions across a wider range of skin diseases.

Metrics Breakdown:

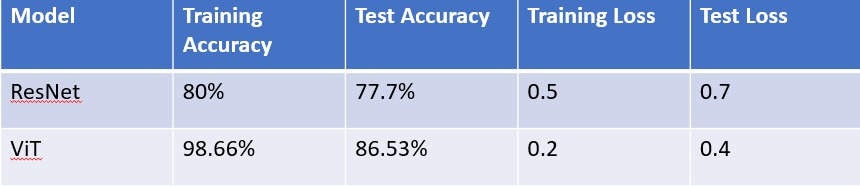
* Precision: The ViT model achieved high precision for most classes, meaning it was successful in correctly identifying positive cases without a high rate of false positives.
* Recall: The model also showed strong recall, particularly for diseases like melanoma, which is critical for early detection in real-world clinical applications.
* F1-Score: The F1-score, which balances precision and recall, was also improved, reflecting a more balanced classification across all 14 skin disease classes.

Advantages of ViT:

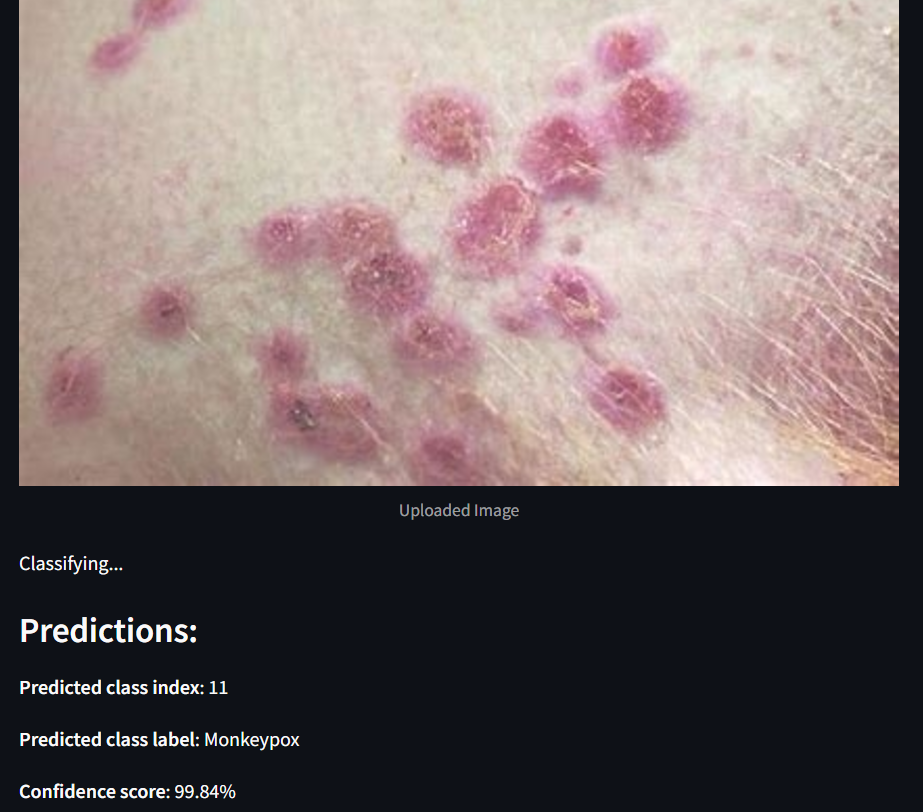
* Global Attention Mechanism: Unlike CNNs, ViT uses self-attention mechanisms to capture global dependencies between image patches. This enables the model to understand spatial relationships across the entire image, which is crucial for classifying complex skin lesions that might be subtle or have varying textures.
* Better Generalization: The ViT model's ability to generalize better to the test data resulted in improved performance, as evidenced by the higher test accuracy.

#### 

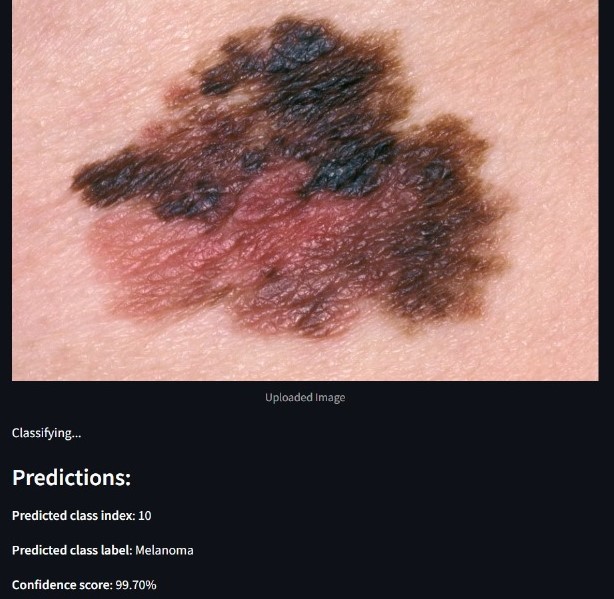
#### **4.5 Comparison of ResNet and ViT Models**



* ResNet vs. ViT: As seen in the table, the ViT model outperformed ResNet on both the training and test sets. The ViT's higher training accuracy indicates better learning of the features, while its higher test accuracy suggests improved generalization to unseen data.
* Loss Comparison: The ViT model also had a lower test loss, indicating that it made fewer errors in predicting the correct labels for the test set.

**4.6 Deployment Results :**





#### 

#### **4.6 Discussion**

The results clearly demonstrate that Vision Transformer (ViT) outperforms ResNet (CNN) in skin disease classification, achieving a higher test accuracy and better generalization. The key reasons for this improvement include:

* Global Context Understanding: ViT’s ability to process the entire image through its attention mechanism allows it to capture complex spatial relationships that are crucial for accurately identifying skin lesions.
* Scalability: ViT scales well with large datasets, and fine-tuning a pre-trained model on our dataset allowed us to achieve high performance even with limited training data.
* Robustness: The ViT model was more robust to variations in skin lesions, which is a common challenge in medical imaging tasks where lesions may vary in size, texture, and appearance.

Despite the improvements, ViT still faces challenges in certain areas:

* Class Imbalance: Although we used a WeightedRandomSampler to address class imbalance, some rare skin diseases still had lower classification performance. Future work could involve additional data augmentation or model fine-tuning to improve performance on underrepresented classes.
* Complexity: ViT’s architecture is more complex and computationally intensive compared to CNNs, making it more resource-demanding during both training and inference.

The Vision Transformer (ViT) model showed a remarkable improvement in the classification of skin diseases compared to the ResNet-based CNN model. With a test accuracy of 86.53%, the ViT model demonstrated its potential for real-world applications in skin disease detection, particularly for challenging conditions such as melanoma. While ViT outperformed CNN models, there is still room for further optimization, such as incorporating advanced augmentation techniques and exploring other deep learning architectures. This work contributes to the growing body of research on using transformer-based models in medical imaging, showing that ViT holds great promise for automated dermatology systems.

**Chapter 5**

**Conclusion**

This project aimed to develop an automated system for the classification of skin diseases using deep learning models. Specifically, the focus was on comparing the performance of Convolutional Neural Networks (CNNs), using ResNet, and Vision Transformers (ViT), with the goal of identifying skin lesions accurately for better diagnostic assistance in healthcare.

#### **5.1 Key Findings**

1. Initial Approach with ResNet:
   * Initially, the project employed a ResNet-based CNN model, which is a widely used architecture for image classification tasks. The ResNet model was trained for 50 epochs, but the results were suboptimal, with a test accuracy of 77.7%. This performance was limited due to the inherent nature of CNNs, which focus on local feature extraction and struggle to capture global relationships in images.
   * Despite achieving reasonable training accuracy, the model’s generalization on unseen data was poor, primarily because of overfitting to the training set and its inability to handle the complexity of skin lesions effectively.
2. Switch to Vision Transformer (ViT):
   * To overcome the limitations of ResNet, the model was switched to a Vision Transformer (ViT), which leverages self-attention mechanisms to capture global dependencies across an image. ViT processes images as sequences of patches, enabling the model to consider the entire image holistically.
   * After fine-tuning the pre-trained ViT-base-patch16-224-in21k model on the skin disease dataset for 10 epochs, the test accuracy improved to 86.53%, significantly outperforming ResNet. The ViT model achieved a training accuracy of 98.66%, indicating that it effectively learned the features from the training set and generalized well to the test data.
3. Improved Generalization and Performance:
   * The ViT model outperformed the ResNet model due to its ability to capture global spatial relationships between image patches, which is crucial for accurately classifying subtle differences between various skin lesions.
   * The ViT model demonstrated better precision, recall, and F1-score across most skin disease classes, particularly those that require precise differentiation, such as melanoma, which is critical for early cancer detection.

#### **5.2 Contributions**

This project contributes to the ongoing research in the application of deep learning models, particularly transformer-based architectures, for medical image analysis. The key contributions include:

* Evaluation of ResNet for Skin Disease Classification: While ResNet is a powerful tool for general image classification tasks, this project highlighted its limitations in capturing global relationships in medical imaging tasks such as skin disease detection.
* Success of Vision Transformer for Skin Disease Classification: The transition to ViT proved that transformer-based models, originally designed for NLP, are highly effective for image classification tasks, especially when global context and long-range dependencies are important. ViT was able to achieve superior performance compared to CNNs, setting a new benchmark for skin disease classification in this project.
* Comprehensive Performance Evaluation: The use of multiple metrics, including accuracy, precision, recall, F1-score, and confusion matrix, allowed for a detailed assessment of the model's performance, ensuring a robust evaluation of the classification system.

#### **5.3 Challenges and Limitations**

Despite the significant improvements with ViT, several challenges and limitations remain:

* Class Imbalance: Although the class balancing technique with WeightedRandomSampler was implemented, some classes still showed lower classification accuracy. This suggests that additional techniques, such as focal loss or oversampling, could further improve performance on underrepresented classes.
* Computational Complexity: ViT models are more computationally intensive compared to CNNs due to their reliance on the self-attention mechanism and large numbers of parameters. This results in longer training times and higher memory requirements. Future work could explore optimizations or more lightweight transformer models to make ViT more feasible for real-time applications.
* Data Quality and Size: While the dataset used in this project was sufficient for demonstrating the effectiveness of ViT, the accuracy could be further improved with larger and more diverse datasets. Moreover, additional data augmentation techniques could help the model generalize better to unseen variations in skin lesions.

#### **5.4 Future Work**

The following steps could be considered for future work to further improve the system’s performance and extend its capabilities:

1. Data Augmentation: Implementing more advanced data augmentation strategies such as random rotations, flipping, color jittering, and zooming can help the model become more robust to variations in image appearance and improve generalization.
2. Advanced Class Balancing: Exploring advanced class balancing techniques like focal loss, SMOTE (Synthetic Minority Over-sampling Technique), or oversampling of underrepresented classes can help improve performance on rare skin diseases.
3. Real-time Deployment: Moving forward, the model can be integrated into a real-time skin disease detection system for practical use by healthcare professionals, especially in remote areas where dermatology expertise is limited. A mobile or web-based application could be developed for easy access and use by non-experts.
4. Integration with Telemedicine: This model could also be integrated with telemedicine platforms for remote diagnosis, allowing patients to upload images of their skin lesions and receive an initial classification of their condition. This would enhance accessibility to dermatological services in underserved areas.
5. Fine-Tuning ViT: Future research can focus on further fine-tuning the Vision Transformer model with more diverse datasets or using hybrid models that combine the best of both CNNs and transformers to optimize performance.
6. Ensemble Methods: Implementing an ensemble of models (e.g., combining ResNet with ViT) could be another avenue for improving accuracy and robustness, particularly in difficult cases where single models may struggle.

#### **5.5 Final Conclusion**

In conclusion, the transition from the ResNet CNN model to the Vision Transformer (ViT) model led to significant improvements in the classification of skin diseases. The ViT model not only outperformed ResNet in terms of accuracy but also demonstrated its potential for real-world applications in automated skin disease detection. This project contributes to the growing field of deep learning in medical imaging, particularly in dermatology, and showcases the ability of transformer-based architectures to handle complex, high-dimensional data like skin lesions. By exploring further optimizations, the model can be improved and potentially deployed in clinical settings for real-time diagnosis and decision support.

**Chapter 6**

**References**

1. Advances in Diagnosis of Skin Lesions Using Dermoscopic Images  
[*https://ieeexplore.ieee.org/document/9858890*](https://ieeexplore.ieee.org/document/9858890)  
2. Deep Learning in Skin Disease Image Recognition: A Review  
[*https://ieeexplore.ieee.org/document/9256314*](https://ieeexplore.ieee.org/document/9256314)  
3. Deep Learning-Based Dermatological Condition Detection  
[*https://ieeexplore.ieee.org/document/10342692*](https://ieeexplore.ieee.org/document/10342692)4. IDENTIFICATION OF SKIN DISEASES FOR VARIOUS SKIN TYPES  
[*https://www.researchgate.net/publication/377260656\_Machine\_Learning\_Study\_Identification\_of\_Skin\_Diseases\_for\_Various\_Skin\_Types\_Using\_Image\_Classification*](https://www.researchgate.net/publication/377260656_Machine_Learning_Study_Identification_of_Skin_Diseases_for_Various_Skin_Types_Using_Image_Classification)5. Expert System for Diagnosis Fever with Rash  
[*https://ieeexplore.ieee.org/document/10107936*](https://ieeexplore.ieee.org/document/10107936)  
6. Skin Lesion Segmentation Based on Vision Transformers and Convolutional Neural Network—A Comparative Study  
 [*https://www.mdpi.com/2076-3417/12/12/5990*](https://www.mdpi.com/2076-3417/12/12/5990)  
7. All those exanthematous fevers are not COVID-19  
 [*https://www.qscience.com/content/journals/10.5339/avi.2021.1?crawler=true*](https://www.qscience.com/content/journals/10.5339/avi.2021.1?crawler=true)

8. . Machine learning approach for classification of maculopapular and vesicular rashes