An Approach to Improve the Efficiency of Deep Learning Framework for Skin Cancer Classification Utilizing Dermoscopy Images

By
Md. Abdul Rabbi Rahat
ID: 213-15-4583

Mehedi Hasan Salman
ID: 213-15-4523

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Dr. Md. Fokhray Hossain
Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 17, 2025

APPROVAL

This Project titled **"An Approach to Improve the Efficiency of Deep Learning Framework for Skin Cancer Classification Utilizing Dermoscopy Images"**, submitted by **Md. Abdul Rabbi Rahat**, ID No: **213-15-4583** and **Mehedi Hasan Salman,** ID No: **213-15-4523** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **17 September, 2025**.

BOARD OF EXAMINERS

Dr. Bimal Chandra Das — **Chairman**
Professor and Dean (in-charge)
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Shah Md Tanvir Siddiquee — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Hasanuzzaman Dipu — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Nazibur Rahman — **External Examiner**
Technical Lead – Database Administrator
Telenor – Grameen Phone Account

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Md. Fokhray Hossain**, **Professor**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Hossain

Dr. Md. Fokhray Hossain
Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Md. Abdul Rabbi Rahat
Student ID: 213-15-4583
Department of Computer Science and Engineering
Daffodil International University

Salman

Mehedi Hasan Salman
Student ID: 213-15-4523
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Dr. Md. Fokhray Hossain, Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Deep Learning** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Accurate identification and classification of skin cancer play an important role in early diagnosis, which is vital for reducing the mortality rate worldwide. However, challenges such as the variation in the appearances of skin lesions, class imbalance and limited computing resources make it hard to develop reliable models. Consequently, in this study, we propose a novel deep learning framework for skin cancer classification combined with self-supervised learning (SSL), multi-axis attention mechanisms and knowledge distillation to enhance the accuracy and efficiency. In order to overcome the dataset imbalance, medical-aware augmentation and dual-level class balancing methods such as focal loss and class weighting, were applied during training. Based on SSL, our framework utilizes MoCo-v3 for robust feature extraction from unlabeled data that helps the teacher model using a Vision Transformer (ViT) backbone enhanced with Low Rank Adaptation (LoRA) and multi-axis attention mechanism to identify the complex pattern of skin lesions. Knowledge distillation transfers the knowledge from the teacher model to a lightweight student model based on TinyViT with custom modifications that achieve 94.14% accuracy on the PAD-UFES-20 dataset in an efficient manner using only 5 million parameters. Evaluated on benchmark datasets including HAM10000, ISIC-2019 and Pad-UFES-20, the teacher model with an accuracy of 93.72%, 92.86% and 94.56%, respectively, outperformed both ensemble and pre-trained baseline models such as ConvNeXt, ResNet101, DenseNet201, EfficientNetB4, TinyVit, EfficientNetB1, MobileNetV3, ResNet18, DenseNet121 and ViT with a significant improvement in accuracy, precision, recall and F1-score. The student model ensured similar performance with great efficiency and this made our model suitable for resource-constrained environments. Ablation studies demonstrated the roles of key components, including LoRA, multi-axis attention and knowledge distillation, whilst explainable AI techniques ensured the attention to clinically relevant features. This research contributes to the detection of skin cancer with an efficient and accurate deep learning framework for clinical implementation in low-resource settings, and opens the door for more medical imaging applications.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter gives an introduction to the skin cancer detection research, describing the motivation, problem statement, objectives, methodology, project outcomes, report organization and concluding summary. It provides the basis for understanding the purpose, approach, and expected contributions of the study to improving the early detection of skin cancer.

1.1 Introduction

Skin cancer is one of the common and dangerous disease where abnormal cells grow out of control in the skin and can be caused by too much exposure to ultraviolet rays from the sun or tanning beds [1]. It affects people all over the world and it comes in major types such as melanoma, which is deadliest, and in non-melanoma forms such as squamous cell carcinoma and basal cell carcinoma. In 2025, experts anticipate that there will be more than 2 million new cases of cancer in the United States, with the most common form of the disease being skin cancer and that some around 105,000 new cases of melanoma will be diagnosed, including many early-stage types [2], [3]. Worldwide, skin cancer is the 17th most common cancer with an estimated 330,000 cases of melanoma and almost 60 thousand deaths in 2022 [4], [5]. Without early treatment, it may spread to other body parts and cause severe health problems such as organ damage, disfigurement, or death, and place a huge burden on health care systems [6].

Traditional identification of skin cancer involves manual methods by the medical professionals, which involve initially thoroughly examining the skin and recognizing suspicious markings of any skin using guidelines such as the ABCDE criteria for melanoma, which includes checking asymmetry, irregularity of the borders, variation of colors, large diameter, and changes over time. Clinicians often add to this process with dermoscopy to take a closer look at the skin structure under a microscope or move to biopsies for microscopic evaluation in a lab setting. Despite their merit, these approaches have significant limitations, such as a rapid decline in quality based on the practitioner's experience, potential for missing small, early signs, and the time involved during comprehensive assessment, which may lead to misdiagnosis in clinical practice [7].

Advances in computer power and data availability have allowed machine learning to play an increased role in the diagnosis of skin cancer by automating the review of dermatological images. Initial machine learning systems required considerable manual selection of features, such as texture or hues in images, but were effective in simple classification tasks. In recent years, deep learning-based methods have revolutionized this field by independently deriving complex patterns from massive skin lesion image collections and performing on or better than humans in categorizing skin lesion abnormality [8]. Comprehensive studies suggest that deep learning facilitates the rapid initial screenings and presents potential to help healthcare providers in underserved areas with a lack of expert access [9]. Nevertheless, current automated tools are facing many challenges, such as an imbalanced dataset, a lack of diverse training data results in biases towards skin tones, creating poor results in uncommon groups. These systems also need a high processing power and have limited possibilities in low-resource environments, as well as often being susceptible to real-world variations, such as inconsistent lighting or image quality. Over-fitting of particular data sets adds to the distrust of outputs, resulting in a more robust need for a solution.

In order to overcome these limitations, the present research aims to propose a knowledge distillation framework based on self-supervised learning methodology, low-rank adaptation methodology and multi-scale methods with the goal of developing a system that could classify skin lesions with high classification accuracy and low resource requirements to make it suitable for medical deployment.

1.2 Motivation

Melanoma and other skin cancers are a key health issue worldwide, with more than ten thousand people being diagnosed each year [10], hence the need for early detection to prevent death. This research is inspired by the potential of deep learning in skin cancer identification, with a series of issues including data imbalance, resource constraint and interpretability, which is a major obstacle in clinical based solutions. The focus of the study is on improving the ability to precisely and efficiently identify skin cancer and assists medical professionals to deliver high-quality patient care and appropriate management of the disease in any medical environment.

1.3 Problem Statement

Skin cancer, specifically melanoma, is a major health issue worldwide and mortality is high due to a significant late detection rate. Despite progress in deep learning, there are several challenges in the clinical setting for using models, including overfitting, resource and data limitations, unbalanced data sets, and target analysis. The following points discuss these particular issues, which should be overcome for the effectiveness of such systems.

1. **Imbalanced Datasets:** Some types of skin lesions are underrepresented in dermatological datasets, leading to a class imbalance, biased model

performance, causing the diagnosis of less common conditions to be less accurate.

2. **Challenges in Accurate Classification:** Current methods have problems in accurately differentiating between malignant and benign lesions, leading to false positive and false negative results, and having an effect that can affect decisions of early diagnosis and treatment.

3. **Overfitting in Deep Learning Models:** The complexity of skin lesion images often leads to overfitting of models as they struggle to generalize, thus resulting in poor model performance on new, unseen data.

4. **Resource Constraints for the Deployment:** The implementation of complex deep learning models in clinical settings is limited due to enormous computational requirements and the need for more efficient models that fit in the hardware constraints of available systems.

1.4 Objectives

The main objective of this research is to develop an efficient and accurate deep learning model for the classification of skin lesions with a self-supervised training approach, knowledge distillation and Multi-Axis Attention mechanism with LoRA. Based on achieving the main objective, this study aims to fulfill the following goals:

- **Address Class Imbalance:** Develop and evaluate strategies such as focal loss and class weighting strategies to address the class imbalance problems in skin lesion datasets, and to advance performance of models for under-represented types of lesions.

- **Improve Model Accuracy:** Enhance the accuracy of skin lesion classification models by taking advantage of the hybrid approach, which includes ViT in combination with multi-axis attention, the methods of using LoRA and self-supervised learning.

- **Mitigate Overfitting:** Use techniques like curriculum learning and adaptive weighting to address overfitting and increase the model's generalization ability to unseen data.

- **Advance Knowledge Distillation Techniques:** Investigate and advance the knowledge distillation technique by considering both soft predictions and feature representations, in order for the student model to effectively imitate the teacher's performance.

- **Optimize Computational Efficiency:** Develop a lightweight student model for efficient use in resource constrained environment with no performance degradation.

Through these objectives, the aim of this research is the development of automated skin lesion classification systems which will, finally, be useful for early detection and also for the whole diagnostic process in clinical practice. This work is also awesome to just expand my knowledge of machine learning and medical imaging.

1.5 Methodology

The methodology to develop an efficient framework for skin cancer classification is based on a systematic approach with the goal of ensuring accuracy and feasibility for practical applications in medical environment. The methodology includes following approaches:

- **Dataset:** Utilize three datasets known as HAM10000, ISIC-2019 & PAD-UFES-20, consisting of over 37,000 images with different categories of lesions and imaging modalities for effective training.

- **Image Preprocessing:** Images are reduced to 224x224 pixels and normalized with a scale of from different pixel sizes of three datasets. Additional steps are taken, such as hair removal using the application of Dull Razor Algorithm and contrast improvement by techniques like CLAHE to provide the input data for the medical imaging task clean.

- **Augmentation:** Implements a medical aware augmentation using light, medium and strong augmentation, such as flips, rotations, color variations and artifacts (noise and blurs) to make models more robust.

- **Class Balancing:** Using the strategy of a dual-level approach with frequency-based weights and better focal loss to address the imbalance in a way that doesn't result in overfitting by focusing on rare classes such as melanoma.

- **Model Selection:** ViT-B/16 as the teacher model for capturing context and TinyViT-5M as the student for its efficiency by the initial feature learning provided by MoCo-v3.

- **Proposed Framework:** It includes the combination of self-supervised learning, knowledge distillation with Multi-Axis Attention and LoRA, protocol for curriculum-based training for optimized performance and lesser resource requirements.

- **Experimental Setup:** Includes uses of Nvidia GeForce RTX 3060 GPU, Intel Core i7-12700, Python 3.8, PyTorch, Albumentations, Conda, & Jupyter Notebooks.

- **Training and Evaluation:** Models are trained with AdamW optimizer, cosine annealing schedulers and using mixed precision for efficiency. Evaluation made using the metrics such as accuracy, precision, recall, F1-score, confusion matrices and ROC with AUC score.

1.6 Project Outcome

Expected outcomes from this research include:

- **Improved Classification Accuracy:** The Model is predicted to give high classification accuracy in terms of classification of skin lesions, particularly melanoma identification will ensure reliable results in medical applications.

- **Enhanced Generalization:** Techniques like curriculum learning and adaptive weighting will help the model to better generalize to varying dermatological images in order to reduce model overfitting.

- **Efficient Deployment**: By deploying efficient architectures like TinyViT and LoRA, we can easily deploy this model on devices with limited resources while preserving the model's performance.

- **Addressing Class Imbalance:** Addressing the class imbalance with the use of focal loss and class weights will better detection accuracy for less common skin conditions.

- **Explainability**: The model will offer transparent decision-making processes, aiding clinicians in making a better understanding of and becoming more comfortable using the model's predictions.

1.7 Organization of the Report

The report starts with Chapter 1, Introduction that includes introduction, motivation, problem statement, objectives, methodology, project outcomes, structure of the report, and summary. Chapter 2, Literature Review contains an introduction, related research, a comparison of existing research, gap analysis, and a summary. Chapter 3, Research Methodology, explains the methodology, the proposed approach, the requirements, the data flow, alternate method, project plan, and the allocation of tasks, and a summary. Chapter 4, Proposed Knowledge Distillation Framework, discusses the proposed framework, self-supervised learning, teacher and student models, distillation process, and a summary. Chapter 5, Implementation and Results, includes environment setup, model training, evaluation metrics, comparative analysis, results and discussion with a summary. Compliance, studying societal impact, sustainability, management, engineering problems, and summary are in Chapter 6, Engineering Standards and Design Challenges. Chapter 7, Conclusion, concludes the study with a summary, limitations, future work and references.

1.8 Summary

This chapter provides a general introduction to skin cancer and the current process used to detect it, including the rationale leading to addressing this global health care challenge. The critical issues related to late detection, and limitations related to dataset imbalances and resource constraints, model performance, etc described in the problem statement. It presents the goals of the study, highlighting the development of an efficient classification framework that uses self-supervised learning, knowledge distillation, and Multi-Axis Attention with LoRA. Additionally, it contains the methodology and expected outcomes, whilst having the structure of the report organized to offer a clear basis for the purpose, approach, and potential contribution to improving the early detection of skin cancers in the clinical environment.

Chapter 2

Literature Review

In this chapter, a review of the literature on deep learning methods for skin cancer classification is conducted, which lays the groundwork for the proposed research. It starts with an introduction in which basic background information and an overview of the main categories of related work are provided. The following sections define the related research, draw comparisons between existing works, analyze gaps and provide a summary with conclusions.

2.1 Introduction

The classification of skin cancer has been one of the most important applications of computer vision and deep learning in medical imaging, aimed at helping dermatologists to diagnose skin lesions more quickly and accurately. With the increasing rate of skin cancer occurrence around the world, the need for a low-cost diagnostic method is developing. These methods have shifted from using traditional machine learning to a variety of deep convolutional neural network (CNNs) techniques, such as transformer models, ensemble-based methods and knowledge distillation techniques. Based on the classification of methods, there are broadly four categories of literature in this space:

1. **CNN-Based Deep Learning Architectures:** The ability to get a hierarchical representation of spatial features makes Convolutional Neural Networks (CNNs) the backbone in medical diagnosis based on images.

2. **Transformer and ViT-Based Methods:** Vision Transformer architectures are an alternative to purely convolutional architectures that make use of self-attention to click on at lengthy vary co-dependencies and contextual data of images.

3. **Ensemble and Hybrid Architectures:** Recent ensemble and hybrid models are more robust and reduce the weaknesses of the single models, they provide a synthesis of complementary advantages from other model architectures, such as CNNs and Transformers.

4. **Knowledge Distillation, Semi-Supervised, and Unsupervised Learning:** In addition to pure architectural breakthroughs, training paradigms such as knowledge distillation, semi-supervised and unsupervised learning are tools that ought to address the issues of scarce data, expensive data annotation and model efficiency.

2.2 Related Research

Integration of machine learning, particularly deep learning methods, on skin cancer has emerged as a potentially powerful tool to increase diagnostic accuracy, swiftness, and accessibility, and such an approach could be deployed to better analyze dermoscopic images to aid clinicians in the detection of malignant lesions.

CNN-Based Deep Learning Architectures: A Novel Deep Convolutional Neural Network (N-DCNN) with 3.3M parameters was proposed by Kaur et al. [11] for skin cancer classification on the ISIC 2019 dataset, which contained 2,882 images after removing duplicates from 7,093 benign and 5,789 melanoma images. Their end-to-end system consisted of segmentation using an Atrous Convolutional Neural Network (ACNN), contrast enhancement using a Multiscale Context Aggregation Network (MCAN), hair removal using morphological operations, and data augmentation including rotation, scaling, and translation. The preprocessed N-DCNN reached the accuracy of 93.40%, precision of 93.45%, recall of 94.51% and F1-score of 93.98%, which were superior to the raw data accuracy of 90.92% and F1-score of 91.70%. These significant improvements were found by bias correction using a correlation-based duplicate elimination technique in combination with quality feature extraction. However, several limitations were identified, such as a lack of segmentation in low-resource environments and a lack of clinical validation.

Utilizing the NIST-National Institute of Standards and Technology's HAM10000 and ISIC datasets, Gururaj et al. [12] used transfer learning for classifying 10,015 dermoscopic images into seven skin lesion classes. Their preprocessing included autoencoder-decoder-based lesion segmentation, hair removal by implementing the dull razor technique, and class balancing through oversampling for ResNet50 and undersampling for DenseNet169. Using the train-test split of 80:20, the oversampled ResNet50 variant got 83.0% accuracy and 84.0% F1-score and the undersampled DenseNet169 model performed up to 91.2% accuracy and 91.7% F1-score. Although the stability of the work was enhanced by a dual-sampling strategy specific to each architecture, optimal performance is limited by hyperparameters that were fixed. Furthermore, extensive parameter optimizations are required in the future in order to further increase the predictive accuracy of the model.

To assist dermatologists in early diagnosis, Reza et al. [13] developed a robust deep learning-based approach for skin cancer classification using 57,536 images. The model had an accuracy of 89.3% ± 1.1% in multiclass classification. In a similar study by Ameri et al. [14] on skin cancer identification, the authors applied a convolutional neural network (CNN) for binary classification of benign and malignant classes. There was no manual manipulation of the images regarding segmentation and feature extraction. Their method reached 84% accuracy in the HAM10000 dataset and clearly demonstrates the importance of using simple architectures when considering the unavailability of high computational resources.

In terms of efficiency, Tuncer et al. [15] proposed a lightweight CNN model referred to as TurkerNet for automated skin cancer classification. Their model attained a high accuracy at low computational cost, acceptable for application on a mobile platform. The model consisted of four blocks: input, residual bottleneck, efficient block and output block that used depth-wise separable convolutions to reduce parameters while preserving feature quality. They validated the model on a publicly available skin cancer dataset of benign and malignant classes and achieved 92.12%% testing accuracy. Although it showed a balance between performance and efficiency, which made it suitable for teledermatology, this study did not employ cross-dataset validation or even properly defined per-class evaluation metrics, specifically for imbalanced abnormalities such as melanoma.

In another research on multi-class classification of skin lesions using the ISIC 2019 [16] dataset, Vega-Huerta et al. [17] developed a custom 19-layer CNN. They started with 2,357 images of nine types and grew to 9,000 through image augmentation via rotation, chopping, and scaling to address the issue of class imbalance. Their architecture consisted of normalization, convolutional, pooling and dropout layers and was trained using the Adam optimizer. The model achieved 94% total accuracy, 89.88% precision, 89.55% recall, 89.4% F1-score and 98.69% specificity on classifications such as dermatofibroma and vascular lesions. Specificity was extremely high, thus allowed negative prediction to be reliable, but poor recall of melanoma at 68% are critical issue.

Transformer and ViT-Based Methods: Using the HAM10000 dataset that consists of 10,015 images belonging to seven classes, Islam et al. [18] compared architectures like ConvNeXt V2, SwinV2, ViT, CvT, DenseNet, ResNet, PVTv2, EfficientNet, EfficientFormer, VGG, MobileNet, MobileViTV2, and GoogLeNet in their research study. After standard preprocessing and splits, ConvNeXt V2 clearly performed better than other models with an accuracy of 93.2%, low inference time, and good confusion matrix results, indicating its construction led to local and global feature capture strengths. The benchmarking process was used to select models that could be used in the real world and better utilize outcomes for the patient. For future work, the authors proposed some optimization methods such as class-aware losses to achieve a better balance.

Moving to the pre-trained ViTs, Flosdorf et al. [19] evaluated ViT-L16 and ViT-L32 on HAM10000 and focused on the importance of melanoma due to its clinical severity. With initial preprocessing, ViT-L16 achieved 92.79% accuracy and 56.10% recall, and ViT-L32 attained 91.57% accuracy and 58.54% recall, while outperforming baselines of decision trees, KNN, and simple CNNs. Although the ViTs are computationally expensive, unscaled, and do not have good interpretability, statistical tests show that they could provide the feasibility of ViTs for quick diagnosis in privacy-constrained settings while still dealing with class imbalance and poor diagnosis accuracy. Future versions of the model could include class-aware loss, robust augmentations, or cost-aware training that is also suited for deployment in a resource-constrained setting.

To integrate deep learning into clinical workflows, Zhang et al. [20] proposed DermViT, a modular version of the Vision Transformer (ViT) that consisted of a Dermoscopic Context Pyramid (DCP) to handle multi-scale variability, hierarchical attention to focus on lesions, and feature gating to suppress artifacts. On the ISIC 2019 dataset, it achieved an accuracy of 86.12% using 40% fewer parameters than ViT-Base and succeeded in demonstrating good generalization. One limitation was that lesion progression over time creates the risk of staging discrepancies. Future work might also involve adding a patient's history to monitor changes in the condition over time.

To classify skin cancer, an improved Vision Transformer model named SkinTrans based on self-attention for discriminative feature extraction was developed by Xin et al. [21]. The model was evaluated on a private dataset of 1,016 dermoscopic images from hospital patients and on the HAM10000 dataset. Their model outperformed baseline ViT, Soft Attention Network, ResNet50, MobileNetV2, InceptionV2 and other CNN models, achieving an accuracy of 94.1% on the clinical dataset and 94.3% on HAM10000. The proposed approach used contrastive learning for increased inter-class discrimination and multi-scale feature mining with overlapping sliding windows and multi-scale patch embedding for capturing local and global context. The model did perform well, but it did not have broad clinical validation and was slow for use on high-resolution images.

Using the HAM10000 dataset and smartphone-captured images, Leema et al. [22] proposed a Lightweight Multi-Scale Vision Transformer (LMS-ViT) for real-time skin cancer classification, overcoming the limitations of CNN for fine-scale lesions using the multi-scale attention mechanism on global structures and textural features. The approach used compact feature fusion to enable fast and platform-independent processing suitable for mobile medical applications, allowing camera-based diagnostics over a smartphone. Experiments showed an accuracy of 90%, 18% higher than the accuracy of traditional CNNs, with 30% lower computing cost and increased precision, recall, and F1-scores. The authors suggested that future studies could use more sensors for multi-modal analysis on more lesion types.

Ensemble and Hybrid Architectures: Utilizing the imbalanced ISIC 2019 and HAM10000 [23] datasets, Pacal et al. [24] proposed a MetaFormer-based model with a CAFormer structure using focal self-attention mechanisms for skin cancer classification. By combining the best of CNN and ViT for optimized feature extraction capabilities, the lightweight model (35.01M parameters) benefits from the state-of-the-art data augmentation, normalization, and scaling methods to deal with the class imbalance. On ISIC 2019, it achieved 92.54% accuracy, 90.41% precision, 87.68% recall, and 88.86% F1-score. On HAM10000, it achieved 95.01% accuracy, 94.70% precision, 92.11% recall, and 93.34% F1-score. There was no clinical metadata integration, but the imbalance limited minority class performance.

Ozdemir et al. [25] proposed a solution to the imbalanced dermoscopic skin lesion classification task on 2,533 images in eight classes using the ISIC 2019 [16] dataset. For fine-grained local texture as well as global context extraction, they developed a hybrid deep framework that integrated ConvNeXtV2 convolutional blocks with focal self-attention modules. The model outperformed standard CNN and ViT baselines, reaching 93.60% accuracy, 91.69% precision, 90.05% recall, and an F1 score of 90.73% with no heavy preprocessing except for normalizing and scaling. Sensitivity to minority classes, as represented by Actinic Keratosis (AK) recall of 78.16% and Squamous Cell Carcinoma (SCC) F1-score of 84.38%, was undermined by class imbalance inherent in the data, and clinical suitability was not evaluated using external validation.

Combining Vision Transformers (ViT-base, ViT-large) and CNN models, including ResNet50, VGG19, DenseNet201, and MobileNetV2, with Swin Transformers Dagnaw et al. [26] proposed a hybrid skin cancer classification model. They evaluated the model on the ISIC dataset, which contains 3,297 dermoscopic images. Preprocessing was done by balancing the cancer class with SMOTE as well as resizing to 224x224 pixels. ResNet50 achieved the highest accuracy of 88.8% followed by 88.6 for the ViT models. Interpretability was improved with Grad-CAM, Grad-CAM++, and Score-CAM. However, the small sample size, the binary classification nature, and the lack of advanced pre-processing such as a hair removal method, all limited the generalizability. Similarly, Tembhurne et al. [27] developed an ensemble-based machine learning (ML) and deep learning (DL) model from the same dataset. After the step of feature extraction and classification, the model achieved an accuracy of 93% on the ISIC Kaggle dataset.

Knowledge Distillation, Semi-Supervised, and Unsupervised Learning: Working on the HAM10000 with 10,015 images and 7 classes, and the ISIC 2019 [28] with 25,331 images and 8 classes, Saha et al. [29] developed a method for skin cancer classification. Both datasets were severely unbalanced, yet they were balanced to 1,200 images per class using augmentation, including shear, rotation, zoom, brightness, and flip. As part of preprocessing, they normalized the images to [0,1] and resized the images to 224x224 pixels. Their method used pretrained teacher models, ResNet50, DenseNet169, VGG16, and VGG19 to supervise lightweight student models using knowledge distillation. With strong multi-class performance statistically validated, student models attained 88.69%-93.24% and 82.14%-84.13%

accuracy on HAM10000 and ISIC 2019, respectively, during Phase 1, and 88.63%-88.89% and 81.39%-83.42% during Phase 2. Preprocessing included no sophisticated cleaning such as hair removal, and student models were still a little worse than teachers. Future work included an exploratory study of multi-teacher distillation, hyperparameter optimization, hair removal, and segmentation.

Liu et al. [30] proposed a relation-aware mutual learning approach for the lightweight classification of skin lesions based on the ISIC 2019 [28] dataset, which contains 25,331 individual dermoscopic skin images belonging to eight imbalanced classes. To overcome class imbalance, they applied weighted cross-entropy loss and performed data augmentation, including flip, brightness, contrast, saturation, and jitter adjustments, adding random noise and mixup to augment our training datasets. The student models, MobileNetV2 or EfficientNetB0, included a sample relationship module (BatchFormer) based on a transformer to encapsulate inter-sample dependencies with mutual learning implemented using Kullback-Leibler divergence, which is switched to inference to enable online deployment on wearable devices. The proposed framework achieved superior results on the validation and test sets: 88.3% accuracy, 85.3% balanced accuracy, 0.980 AUC, and 0.801 mAP with MobileNetV2, and 88.5% accuracy, 87.2% precision-balanced accuracy, 0.981 AUC, and 0.796 mAP with EfficientNetB0.

Morales et al. [31] utilized a heterogeneous dataset of 1,449 labeled and 2,508 unlabeled dermoscopic images from Virgen del Rocio University Hospital, Polesie [32], ISIC, and Kawahara [33] databases to develop an ensemble-based Knowledge Distillation model using semi-supervised CNN models-DenseNet121, ResNet50, and VGG16 for melanoma Breslow Thickness (BT) classification. Student models achieved the best balanced accuracy of 0.7831 ± 0.1225 and AUC of 0.8531 ± 0.0921 on ISIC for binary BT classification after teacher models were refined with labeled data and pseudo-annotated unlabeled images through majority voting. Grad-CAM++ identified regions important for the prediction that led to an improved understanding of the images. Although their proposed approach increased the generalizability, it had high computational costs and the multiclass performance was limited by the similarity between classes.

Using the ISIC 2019 [34] with 23,400 images across six classes, excluding DF and VASC, and HAM10000 with 5,000 images across five classes, Rahman et al. [35] presented an unsupervised k-means method assisted by histogram-based feature extraction and ESRGAN-based super-resolution. After resizing and augmentation of the lesions to 224x224 pixels, they calculated 8x8x8-signed RGB histograms normalized by mean-variance normalization and applied PCA+t-SNE for the dimensionality reduction. Their pipeline outperformed other clustering techniques, achieving 87.77% accuracy on ISIC 2019 and 90.50% accuracy on HAM10000. Their label-free approach reduced the cost of annotation and provided greater image clarity through the application of an optimized ESRGAN. The limitations of this method were the class overlap and t-SNE as well as the high computational cost of ESRGAN.

2.3 Comparison Between Existing Research

In this section, a comparative analysis of the existing studies for the skin cancer classification is provided in terms of the methodologies used, highest reported accuracies with associated algorithms and the years of publication to visualize the changes and performance trends over time in the field.

Table 2.3: Comparative Analysis of Previous Work

Author (s)	Methodology	Best Accuracy	Year
Kaur et al. [11]	End-to-end system with segmentation (ACNN), contrast enhancement (MCAN), hair removal, and data augmentation for N-DCNN	93.40% (N-DCNN)	2025
Gururaj et al. [12]	Transfer learning with lesion segmentation, hair removal, and class balancing using oversampling, undersampling for ResNet50 and DenseNet169	91.2% (DenseNet169)	2023
Reza et al. [13]	Robust deep learning approach for multiclass classification	89.3% (Inception-ResNet-v2)	2022
Ameri et al. [14]	Simple CNN for binary classification without manual segmentation or feature extraction	84% (Custom CNN)	2020
Tuncer et al. [15]	Lightweight CNN (TurkerNet) with depth-wise separable convolutions for binary classification	92.12% (TurkerNet)	2024
Vega-Huerta et al. [17]	Custom 19-layer CNN with normalization, pooling, dropout, and data augmentation for multi-class classification	94% (Custom CNN)	2025
Islam et al. [18]	Benchmarking of multiple architectures including ConvNeXt V2, SwinV2, ViT, with standard preprocessing	93.2% (ConvNeXt V2)	2024

Flosdorf et al. [19]	Evaluation of pre-trained ViT-L16 and ViT-L32 with initial preprocessing	92.79% (ViT-L16)	2024
Zhang et al. [20]	Modular ViT (DermViT) with Dermoscopic Context Pyramid, hierarchical attention, and feature gating	86.12% (DermViT)	2025
Xin et al. [21]	Improved ViT (SkinTrans) with self-attention, contrastive learning, and multi-scale feature mining	94.3% (SkinTrans)	2022
Leema et al. [22]	Lightweight Multi-Scale ViT (LMS-ViT) with multi-scale attention and compact feature fusion for real-time classification	90% (LMS-ViT)	2025
Pacal et al. [24]	MetaFormer-based model (CAFormer) with focal self-attention, data augmentation, and normalization	95.01% (CAFormer)	2025
Ozdemir et al. [25]	Hybrid framework integrating ConvNeXtV2 with focal self-attention modules and sliding window dual-path architecture	93.60% (Hybrid ConvNeXtV2)	2025
Dagnaw et al. [26]	Hybrid model combining ViTs, CNNs (ResNet50, etc.), and Swin Transformers with SMOTE balancing and interpretability tools	88.8% (ResNet50)	2024
Tembhurne et al. [27]	Ensemble of ML and DL models with feature extraction and classification	93% (Ensemble ML, DL)	2023
Saha et al. [29]	Two-phase knowledge distillation using pre-trained teachers (ResNet50, etc.) to supervise lightweight students with data augmentation	93.24% (Custom Student Model)	2025
Liu et al. [30]	Relation-aware mutual learning between identical student models (MobileNetV2, EfficientNetB0) with weighted loss and augmentation	88.5% (EfficientNetB0)	2024
Rahman et al. [35]	Unsupervised k-means with histogram-based features, ESRGAN super-resolution, PCA+t-SNE dimensionality reduction, and augmentation	90.50% (Unsupervised k-means)	2025

2.4 Gap Analysis

The major gaps of existing research on skin cancer classification that this study aims to fulfill include a unified framework incorporating multiple state-of-the-art techniques like self-supervised learning, attention mechanisms, knowledge distillation, optimized class balancing and model interpretation, which are missing or inconsistently used in earlier literature. Here are some of the major challenges:

1. **Limited adoption of Unified processes:** Most existing work has adopted single models like CNNs, transformers, or knowledge distillation, and hardly combined self-supervised learning, attention mechanisms, and optimized class-balancing together into a unified system, which causes poor robustness and efficiency across different environments of medical images.

2. **Lack of use of self-supervised learning:** Self-supervised or contrastive learning is rarely applied in skin cancer classifications. As a result, the utilization of an unlabeled dataset is unavailable for effective pretraining, which is less developed in feature extraction, especially for small or heterogeneous medical datasets, compared to supervised approaches.

3. **Limited use of knowledge distillation:** Knowledge distillation has the potential to generate lightweight models with performance close to teacher models. However, very few studies apply it to skin cancer classification, and the majority of them lack systematic teacher-student pipelines that can balance efficiency and accuracy for clinical deployment and edge deployment.

4. **Poor Interpretability and Explainability:** Interpretability is rarely on the radar in vision transformers and ensemble-based models, and even where explainable AI like CAM, Grad-CAM, LIME, and SHAP are used in a few instances, there are only limited transparency mechanisms on model outputs, inhibiting clinical trust and actual deployment in real life.

5. **Unresolved issues of class imbalance and minority class performance:** While class imbalance has been recognized, only a few studies reliably apply state-of-the-art methods like class-aware losses or synthetic data generation, resulting in low recall and F1 for minority classes like melanoma, as shown by Vega-Huerta et al. [16], which compromises the reliability of the diagnosis.

6. **Lack of Generalizability:** Many models show poor performance in the case of datasets with different imaging conditions and characteristics, as well as low efficacy for cross-dataset validation, although problems with small sample sizes and binary classifications were pointed out by Dagnaw et al. [25].

7. **Challenges in real-time deployment and computational efficiency:** High computational complexity as reported by Morales et al. [29] and Rahman et al. [32], and a focus away from real-time implementation, as reported by Tuncer et al. [15] and Leema et al. [21], are some of the major challenges.

Table 2.4 shows the summary of the methodological gaps from the literature review with Yes or No, indicating if each methodological gap is present or not. It allows precise identification of gaps that require further research and development of our research contribution.

Table 2.4: Gap Analysis of Previous Work

Authors (s)	Self-Supervised Learning	Attention Mechanism	Knowledge Distillation	Interpretability	Class Balancing
Kaur et al. [11]	No	No	No	No	Yes
Gururaj et al. [12]	No	No	No	No	Yes
Islam et al. [18]	No	Yes	No	No	Yes
Flosdorf et al. [19]	No	Yes	No	No	Yes
Zhang et al. [20]	No	Yes	No	Yes	No
Xin et al. [21]	Yes	Yes	No	No	Yes
Leema et al. [22]	No	Yes	No	No	Yes
Pacal et al. [24]	No	Yes	No	No	Yes
Ozdemir et al. [25]	No	Yes	No	No	Yes
Dagnaw et al. [26]	No	Yes	No	Yes	Yes
Tembhurne et al. [27]	No	No	No	No	Yes
Saha et al. [29]	No	No	Yes	No	Yes
Liu et al. [30]	No	Yes	Yes	No	Yes
Morales et al. [31]	Yes	No	Yes	Yes	No
Rahman et al. [35]	Yes	No	No	No	Yes

2.5 Summary

In this chapter, a comprehensive review of the current scene in skin cancer classification with DL methods has been presented, and it has given a strong basis for the proposed research. This study shows the importance of detection of skin cancer and covers the evolution from traditional ML models to advanced deep learning techniques, organizing related research into four main tracks: CNN-based systems, Transformer and ViT-based models, ensemble and hybrid, and knowledge distillation integrated with semi-supervised models. Regarding the related research work, there are some important contributions like Kaur et al. [11]'s N-DCNN, Xin et al. [21]'s SkinTrans, and Pacal et al. [24]'s CAFormer that achieved the highest accuracy of 95.01% among all other reviewed research, despite the difficulty of class imbalance and limited clinical validation. The analysis of related works found an evolution of methodologies from 2020 to 2025, whereas the gap analysis identified the major lack of integrated advanced techniques, limited use of self-supervised learning, lack of interpretability, lack of real-time deployment, etc. These gaps demonstrate a need for a global framework that incorporates several strategies to simultaneously increase robustness, efficiency and clinical applicability, which is the focus of future chapters in this study.

Chapter 3
Research Methodology

This chapter presents the research methodology for the development of the efficient skin lesion classification system, including the methodology overview, describing the details of the dataset, preprocessing, class imbalance approaches, the model selection process, the requirements, dataflow, detailed design and the project plan and the distribution of the task along with a summary.

3.1 Methodology

3.1.1 Overview

In this study, we proposed a comprehensive overview of the research methodology, including the key steps involved in data preparation, model development, and evaluating the model's performance. The methodology exploits state-of-the-art deep learning methodologies to solve the complexities of skin cancer classification by first gathering and processing the dermoscopic images from various data sets in order to promote a wide variety of lesion types and imaging modes. It uses self-supervised learning methods to learn meaningful feature representations from unlabeled data, knowledge distillation to train a small student model under a strong teacher model, and methods to address class imbalance, which contribute to a reliable and effective classification model specifically for medical applications.

3.1.2 Proposed Methodology

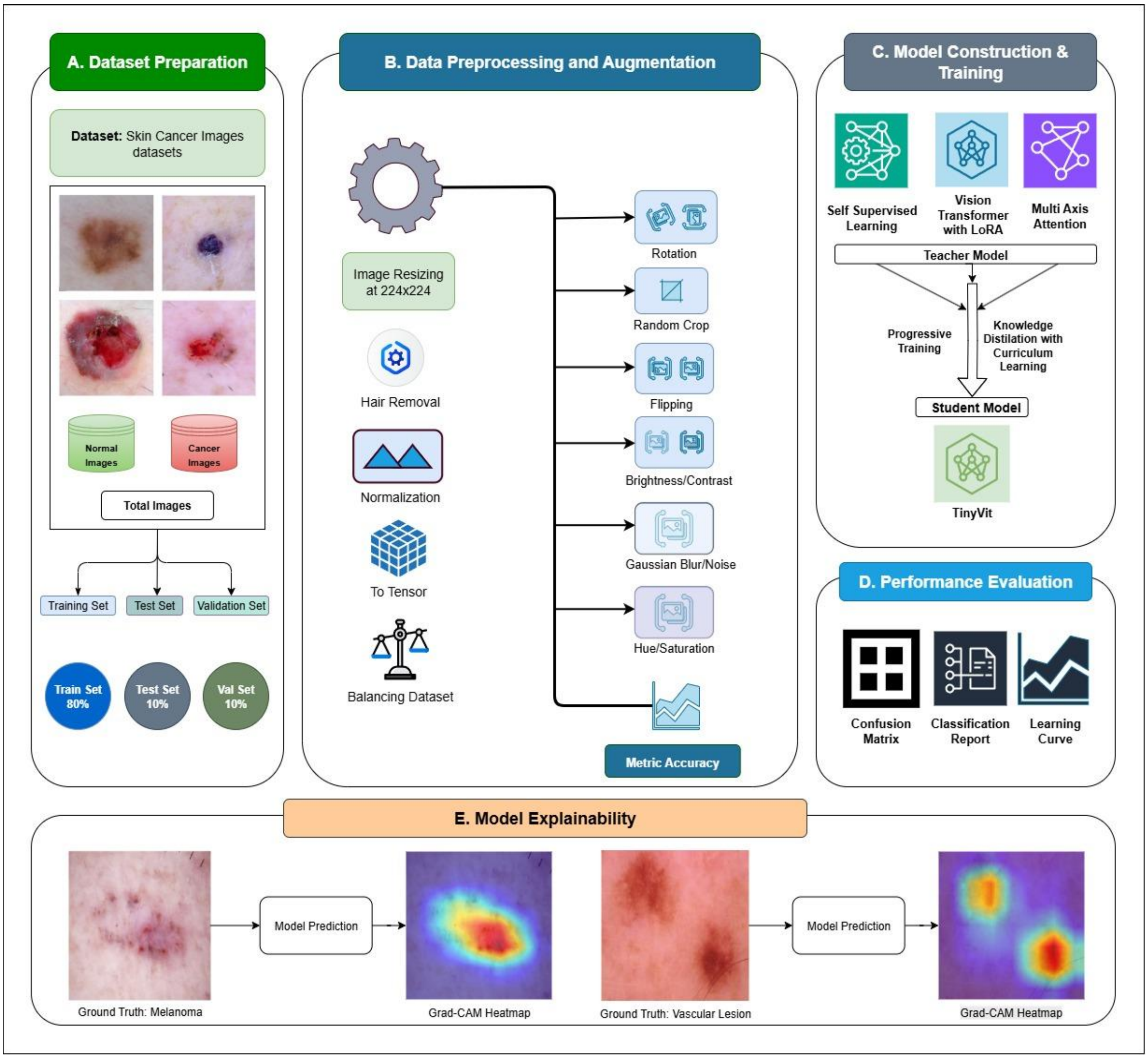


Figure 3.1.2: Proposed Methodology

3.1.2.1 Dataset Description

In this study, three different datasets are utilized to test the performance of the proposed model in skin cancer classification. These datasets provide wide coverage across multiple types of skin lesions, taken via multiple imaging modalities, and across multiple diagnostic classes. Table 3.1.2.1 represent a overview of the number of images and classes for the datasets.

Table 3.1.2.1: Overview of datasets

Dataset	Number of Images	Number of Classes
HAM10000	10015	7
ISIC-2019	25331	8
PAD-UFES-20	2298	6

HAM10000: The HAM10000 [23] dataset, "Human Against Machine with 10,000 Training images", contains 10,015 dermoscopic images at different resolutions ranging from 450x600 to 1022x767 pixels. It consists of seven diagnostic categories: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF) and vascular lesion (VASC). These high-quality dermoscopic images are accompanied by confirmed histopathological diagnoses, which provide the benchmark for skin lesion classification tasks.

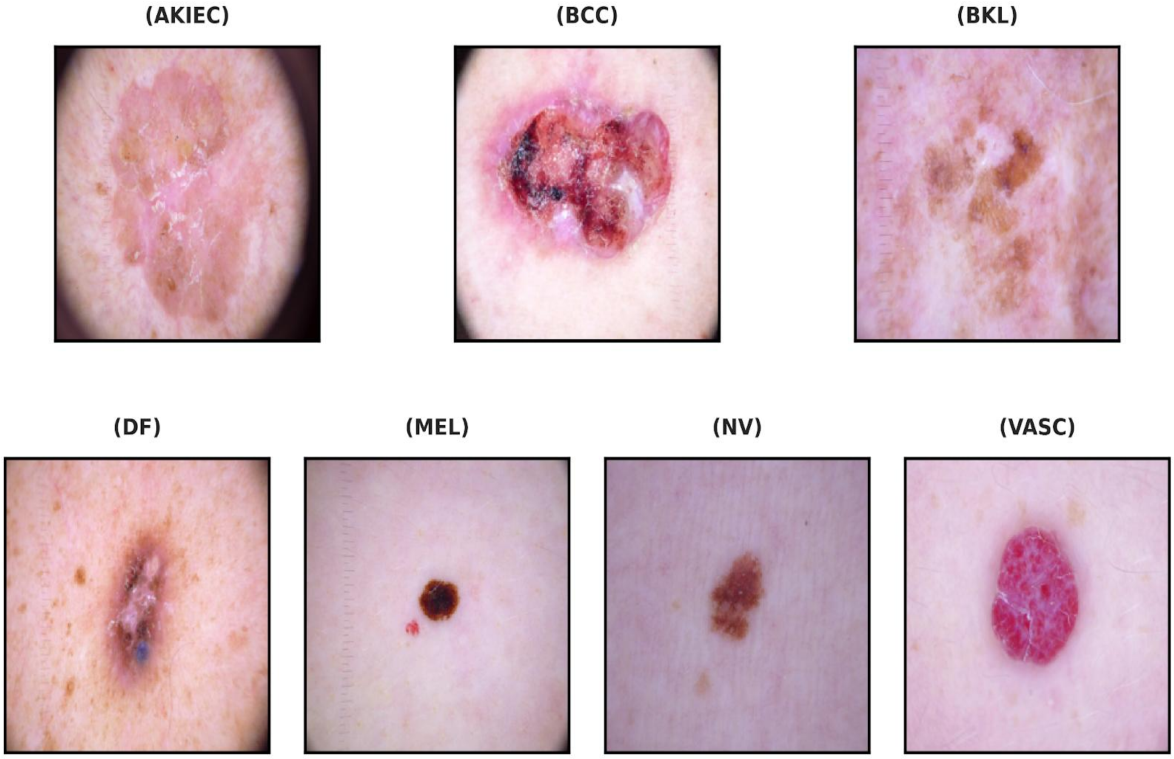


Figure 3.1.2.1.1: Sample Images of each class (HAM10000)

Figure 3.1.2.1.1 illustrates some sample images of each class in the HAM10000 dataset. In terms of key visual characteristics, the pictures reveal: melanomas are irregular, asymmetrical; melanocytic nevi are symmetric and evenly pigmented; basal cell carcinomas are translucent or pearly; actinic keratoses are rough and scaly; benign keratoses are waxy or "stuck on"; dermatofibromas are hard nodules; vascular lesions are either red or purple in color. These morphological details are important for proper classification and are well visualised on dermoscopic imaging.

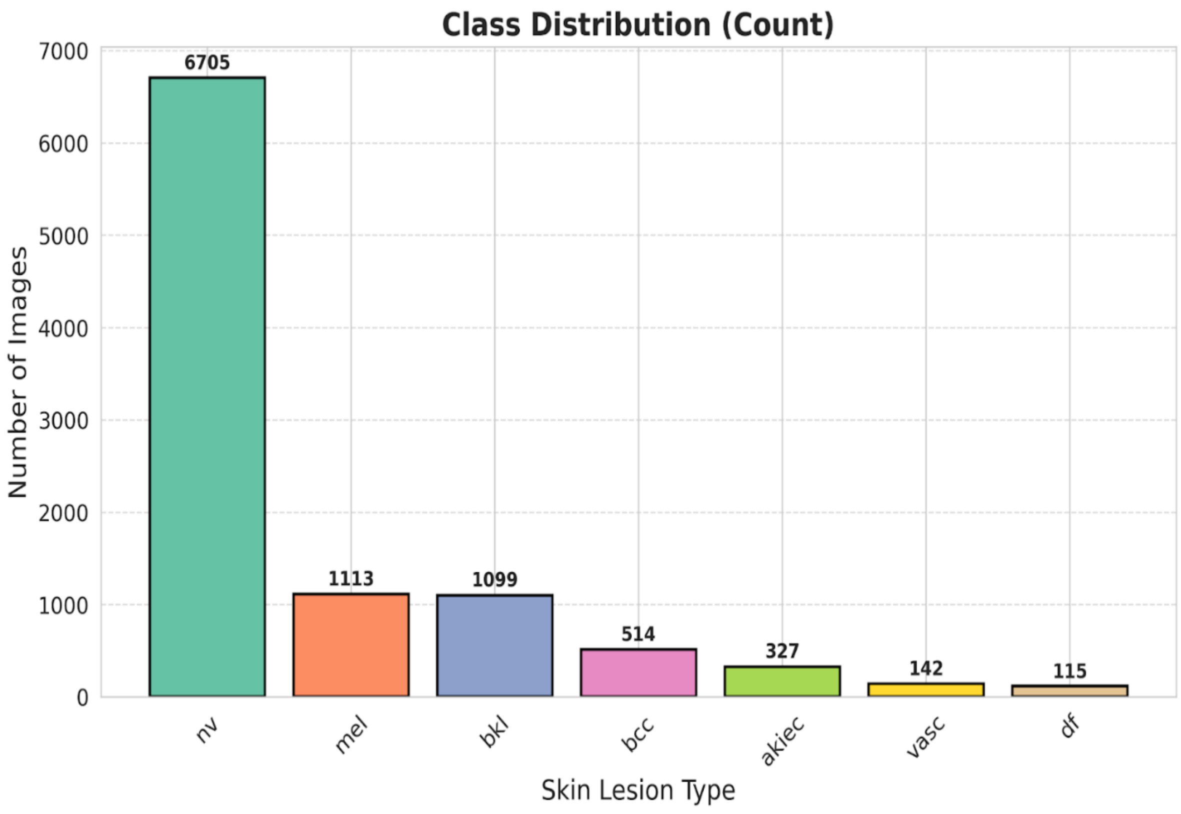


Figure 3.1.2.1.2: Image Distribution of each class (HAM10000)

Figure 3.1.2.1.2 shows the class distribution of the HAM10000 dataset, which shows a significant class imbalance. Melanocytic nevi (NV) represent the largest percentage with 6,705 samples, corresponding to 67% of the total. Other classes contain melanoma (MEL), which includes 1,113 samples, benign keratosis (BKL), which includes 1,099, basal cell carcinoma (BCC), which includes 514, actinic keratosis (AKIEC), which includes 327, vascular lesions (VASC), which includes 142, dermatofibroma (DF), which includes 115. This lack of balance creates difficulties in training the model since the dominant class is significantly more numerous than the others, including the important melanoma category.

ISIC-2019: The ISIC-2019 [36] dataset from the International Skin Imaging Collaboration 2019 Challenge includes 25,331 dermoscopic images and has resolutions usually between 1024x1024 and 2048x1536 pixels. It describes eight diagnostic categories and is a generalization of previous data sets by including squamous cell carcinoma and providing an increased scale for learning. Each image contains metadata describing the patient, lesion location, etc.

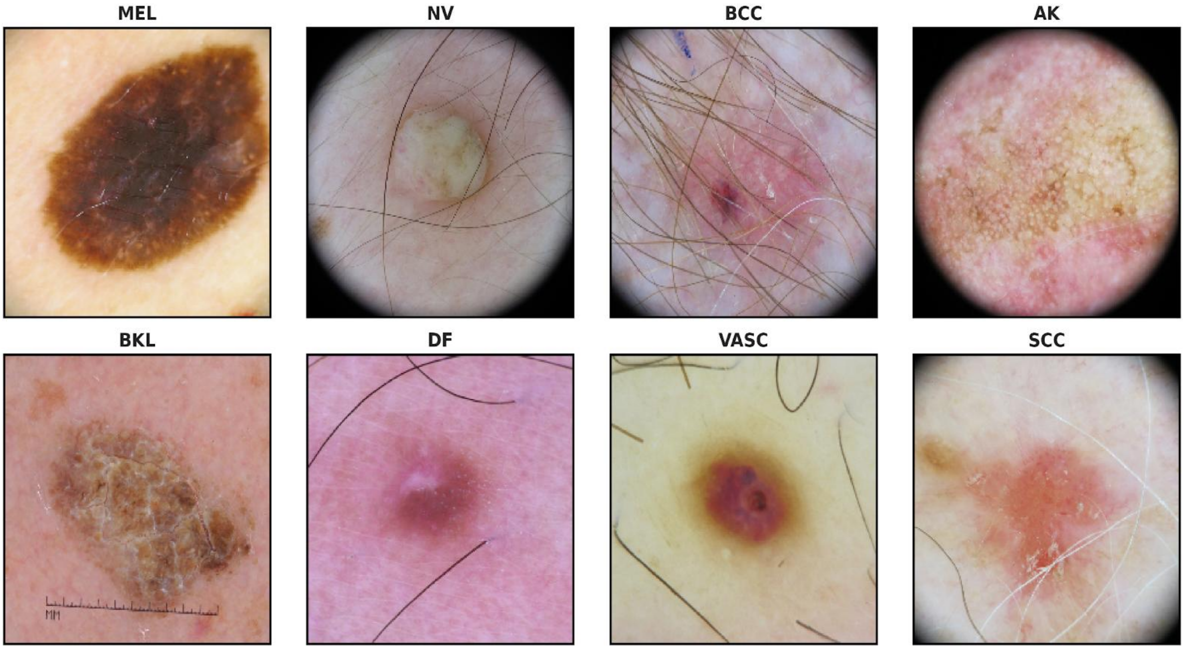


Figure 3.1.2.1.3: Sample Images of each class (ISIC 2019)

Figure 3.1.2.1.3 shows a sample image for each class of the ISIC-2019 dataset, highlighting the general spectrum of lesion types. These high-resolution dermoscopic images demonstrate well-defined borders and surface morphology with the presence of thick, scaly, or crusted features for squamous cell carcinoma. The high-quality imaging criteria are effective for the extraction of features and classification.

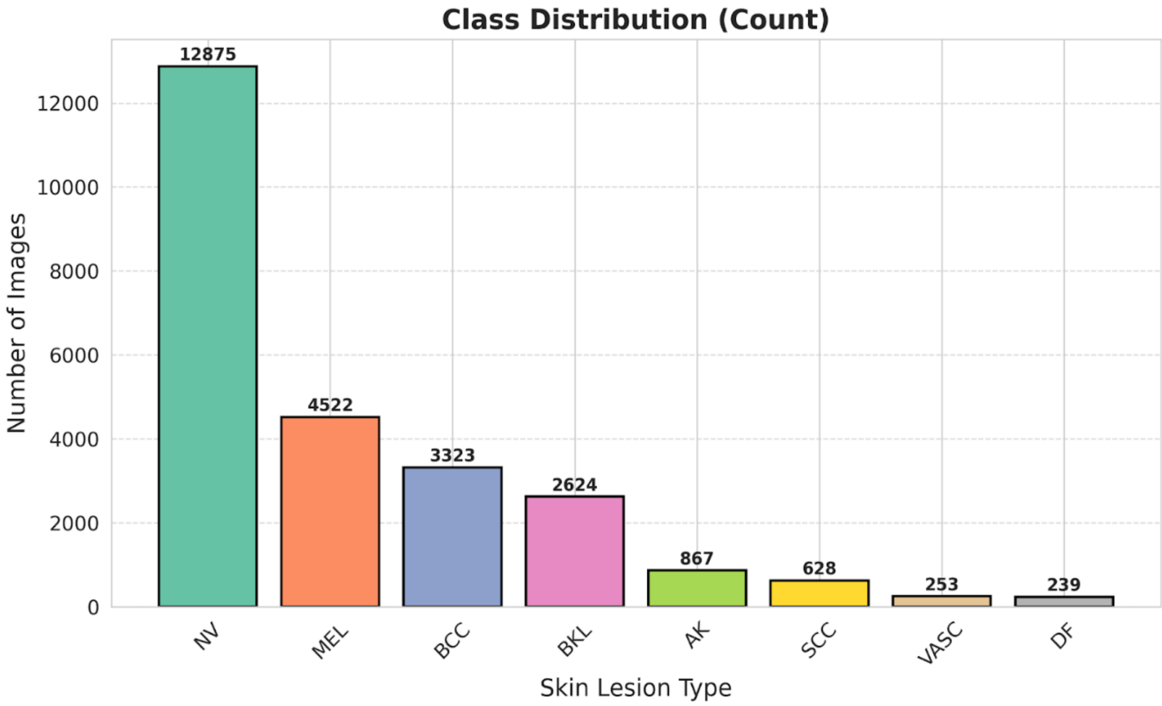


Figure 3.1.2.1.4: Image Distribution of each class (ISIC 2019)

Figure 3.1.2.1.4 illustrates the class distribution on the ISIC-2019 dataset with an even higher imbalance. The largest class of images is NV-12,875, followed by MEL-4,522, BCC-3,323, and BKL-2,624. Smaller case numbers include actinic keratosis AK with 867, SCC with 628, VASC with 253, and DF with 239. The difference between the majority class and the minority class is more than 50-fold, increasing the importance of imbalance handling in model development.

PAD-UFES-20: The PAD-UFES-20 [40] dataset is from the Federal University of Espírito Santo and includes 2298 clinical images ranging in size from 720x480 to 4000x3000 pixels. It contains six diagnostic categories that are captured with smartphones and professional cameras, and it can also be used for simulating the actual environment outside dermoscopy.

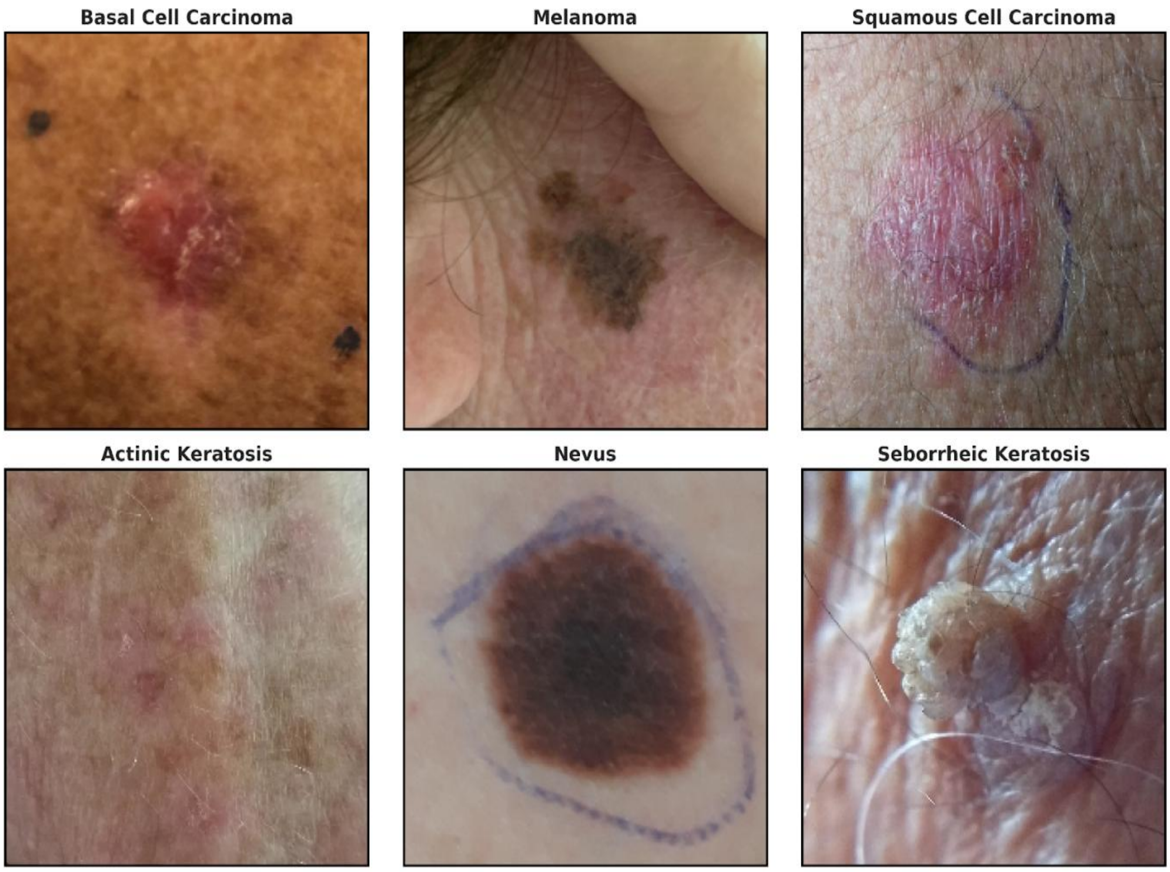


Figure 3.1.2.1.5: Sample Images of each class (PAD-UFES-20)

Figure 3.1.2.1.5 shows example images from each class in the PAD-UFES-20 dataset, which shows the variation in lighting, angles, and image quality. This variability is related to the clinical reality involving different imaging conditions of the lesions while applying dermatological imaging.

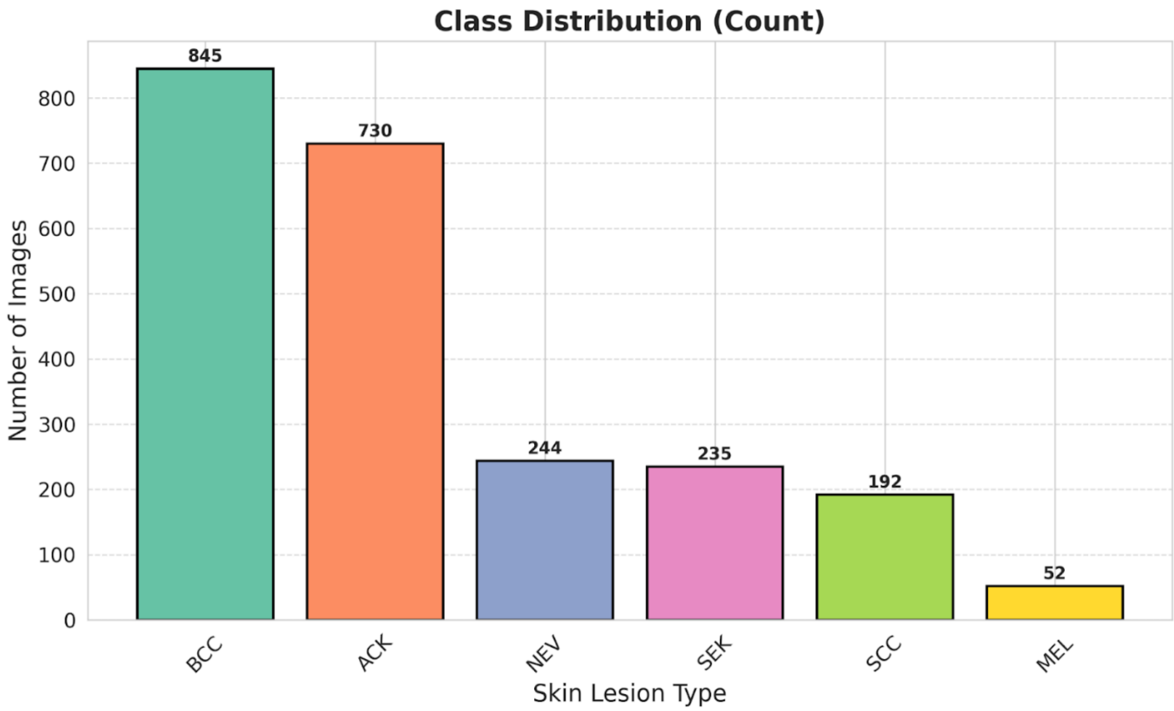


Figure 3.1.2.1.6: Image Distribution of each class (PAD-UFES-20)

As Figure 3.1.2.1.6 indicates, the class distribution in the PAD-UFES-20 dataset resembles this unique imbalance pattern. Basal cell carcinoma (BCC) is the most dominant cancer with 845 images, followed by actinic keratosis (ACK) with 730 images, providing a relative balance at the top. Other classes are nevus (NEV) with 244 images, seborrheic keratosis (SEK) with 235 images, squamous cell carcinoma (SCC) with 192 images, and melanoma (MEL) with only 52 images. The low number for melanoma highlights the need for measures such as additions to compensate for the unbalanced training for this important tumor.

3.1.2.2 Image Preprocessing Techniques

In this section, we describe the pipeline used to preprocess dermoscopic image data for deep learning to ensure that all images in the dataset are clean, standardized and prepared for training while retaining all clinically useful diagnostic features, including lesion boundaries and colors.

Data Loading, Cleaning and Hair Removal:
We load dermoscopic image datasets stored in directories along with associated metadata files that include the diagnosis of the lesions and patient information, respectively. A single dataframe is produced by associating each image with its metadata identifiers to ensure that all the images are associated with their respective labels, generally skin lesion types. These labels are encoded as numeric indices so that the same value is used consistently during training. To ensure the reliability of the dataset, we exclude samples with duplicated instances based on

correlation and metadata, and retain only the unique lesion to avoid bias. For all datasets except PAD-UFES-20, we use a Dull Razor hair removal algorithm based on OpenCV that converts images to RGB, identifies the hair region using morphological filtering, and reconstructs those areas using surrounding texture to create clean images, thereby enhancing the visibility of the lesion for proper classification. The PAD-UFES-20 dataset, composed of clinical photographs instead of dermoscopic images, does not require hair removal since clinical images commonly have fewer hair artefacts because of the lower magnification and wider field of view, minimizing the chances of hair obscuring lesion details important for diagnosis. The cleaning process produces robust and reliable data for further analysis.

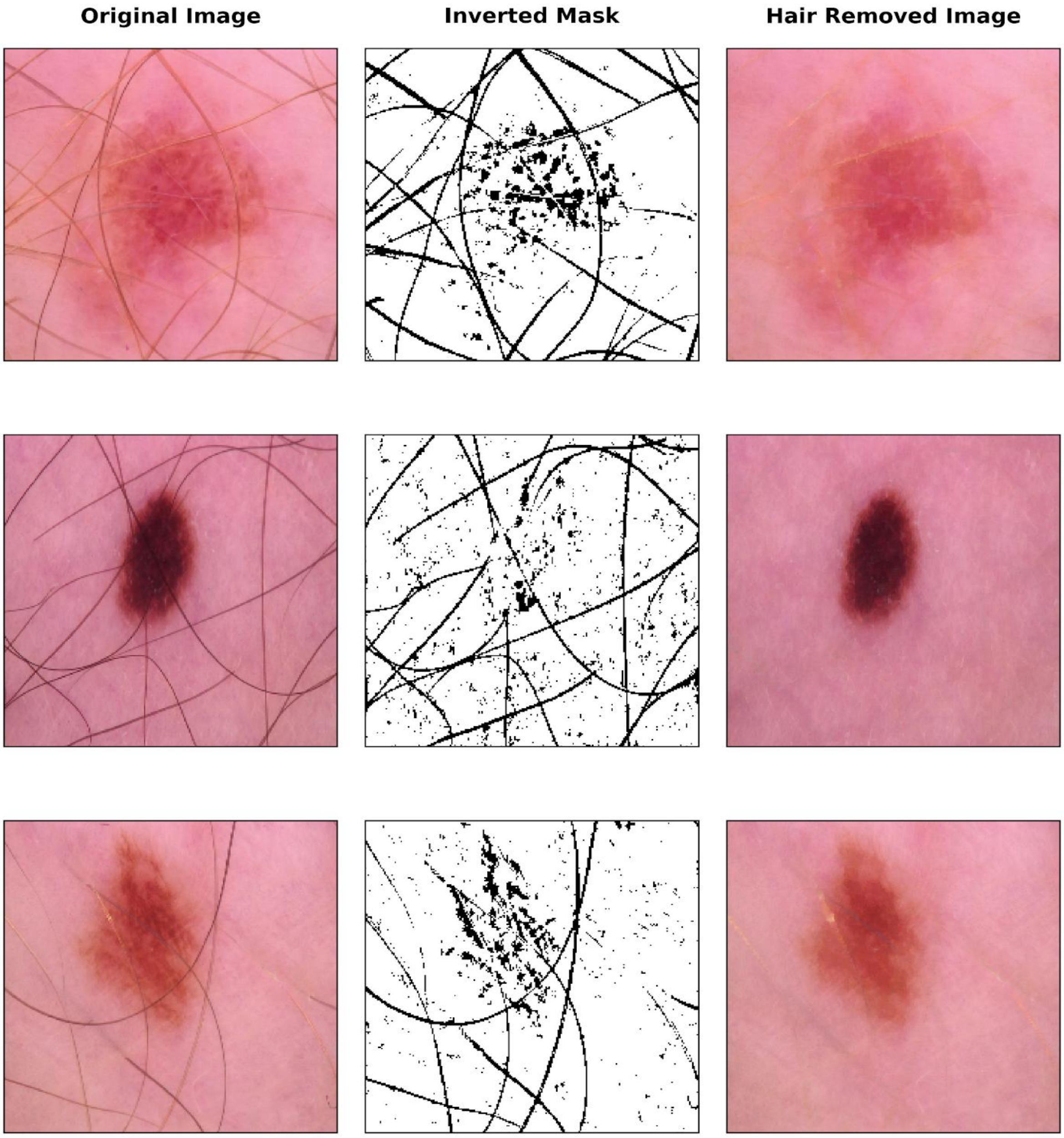


Figure 3.1.2.2: Hair removal representation

Dataset Splitting:

We divide all dermoscopic image data into train, validation, and test sets to facilitate proper model training and evaluation. A stratified approach is employed to maintain class balance across all sets so that each subset represents the general distribution of lesion types. Around 10% of the data is allocated as the test set for the final model evaluation, and the rest is divided into around 80% for training and 10% for the validation set for performance monitoring and to avoid overfitting. This partitions ensure that different datasets such as HAM10000 and PAD-UFES-20 can be well generalized and the variety of lesion type is maintained for reliable evaluation.

Image Standardization and Augmentation:

We standardize images for pretrained deep learning models by resizing images to a consistent resolution, maintaining lesion information while keeping the processing time low for model compatibility and the storage requirements low. We normalize the pixel values to stabilize training and increase performance. To cope with real-world variation, we introduce dermatology-specific augmentations for supervised and self-supervised learning that are necessary for generalization without changing diagnostic features such as lesion boundaries and colors. In particular, we rotate and flip images to simulate various angles of the lesion, change the lighting using brightness, contrast and adaptive histogram equalization, and change the color and saturation of the image to simulate equipment and skin tone artifacts. Geometric transformations, such as scaling, translation, rotation, elastic deformations, and grid distortions are used to preserve lesion boundaries and to simulate perspective variations. Slight noise, blur, motion blur, and dropout patches are introduced in order to simulate faults such as occlusions. Supervised learning with a light augmentation to deal with accurate classification, and self-supervised learning with a heavy augmentation to sample different image pairs for robust feature learning. These procedures are shown to provide good generalization and good diagnostic accuracy.

Data Preparation for Training:

We create a pipeline to efficiently prepare dermoscopic images and their corresponding annotation data for model training, such that the images and their corresponding annotation data can be delivered in a seamless manner during contrastive and supervised learning. For contrastive learning, we generate pairs of image augmentations from each image using the dermatology-specific augmentation process, and train a model that learns across multiple views while preserving the diagnostic features of the images. In supervised learning, every image is associated with its corresponding numerical label, which prepares the data for classification. Such a setting allows for efficient batch processing, which makes it easier for the model to learn effectively from the dataset and supports both kinds of learning schemes.

3.1.2.3 Class Imbalance Mitigation Strategies

This study uses class weighting to cope with the class imbalance in the datasets instead of using sampling techniques. The choice has three reasons. First, synthetic sampling techniques such as SMOTE produce synthetic samples containing synthetic artifacts, which may not reflect realistic clinical features such as subtle textures or color differences among skin lesions. Second, medical datasets are usually small; losing all real samples from the dataset is highly undesirable, undersampling would lose valuable information, and oversampling may overfitting to repetitive patterns. Third, class weighting uses no additional data pre-processing, and the resources can then be allocated to model training. By mathematically changing the loss function, it puts more weight on minority classes without changing the original data. In terms of the solution method, a two-level balancing strategy based on static and dynamic mechanisms is proposed to consider both frequency and difficulty of imbalance.

Level 1: Frequency-Based Static Balancing

This level is based on static weights of clinical severity and frequency. To circumvent this problem, higher weights are applied to rare malignant classes like melanoma and squamous cell carcinoma to make them more influential during training. This corrects the variance towards the majority classes in the medical data set, so that the model gives more importance to clinically relevant cases.

Level 2: Difficulty-Based Dynamic Balancing

This level applies instance-specific weights dynamically through an enhanced focal loss. The focal loss formula is:

$$FL(\rho_0) = -\alpha(1 - \rho_0)^{\gamma}.\log(\rho_0) \qquad (1)$$

where ρ_0 is the predicted probability for the observed class, α is the class imbalance control parameter (between 0 and 1), which increases the prediction of minority classes, and γ is the loss for well-classified samples that is reduced by using a focusing parameter of 0. Larger γ values give greater weight to instances that are difficult to predict, which can improve learning on difficult examples. Both parameters are optimized in order to achieve good performance.

Class Weight Distribution

The weights are trained specifically for the characteristics of each dataset, favouring life-threatening or complex classes, with nevus as the reference (1.0). Table 3.1.2.3 represents the full weight distribution of all classes across three datasets.

Table 3.1.2.3: Class Weight Distribution

Dataset	Class Name	Weight
HAM10000	Melanoma	3.2
	Actinic Keratosis	3.0
	Dermatofibroma	2.5
	Vascular Lesions	2.2
	Basal Cell Carcinoma	1.8
	Benign Keratosis	1.8
	Nevus	1.0
ISIC-2019	Dermatofibroma	4.2
	Vascular Lesions	4.0
	Squamous Cell Carcinoma	3.5
	Actinic Keratosis	3.2
	Melanoma	2.8
	Basal Cell Carcinoma	2.2
	Benign Keratosis	2.0
	Nevus	1.0
PAD-UFES-20	Melanoma	4.5
	Squamous Cell Carcinoma	3.2
	Actinic Keratosis	2.8
	Basal Cell Carcinoma	2.2
	Seborrheic Keratosis	1.8
	Nevus	1.0

3.1.2.4 Model Selection

In this study, we chose some deep learning models to compare their performance in skin cancer classification. Models contain both convolutional and transformer-based architectures selected for demonstrated performance over medical images, class imbalance, and interpretability. Based on previous research, only the models providing the right balance between accuracy, high speed, and applicability to datasets such as HAM10000, ISIC-2019, and PAD-UFES-20 are considered for the selection. Below, each model is presented in an objective manner, summarizing its defining features, strengths when applied to skin lesions, and typical solutions to problems of data imbalance and model interpretability.

Vision Transformer (ViT):

ViT-based skin lesion classification approaches utilize self-attention to capture long-range dependencies and contextual information and perform well on challenging medical imaging applications. Furthermore, they are modular enough to provide hierarchical attention and feature gating mechanisms that are inspired by clinical reasoning but also allow targeted interpretability by leveraging XAI methods like Grad-CAM and Grad-CAM++. ViT models also benefit from large-scale pretraining and transfer learning, leading to fast convergence and strong performance in the presence of visual artifacts and heterogeneity in lesion types.

EfficientNetB4:

EfficientNetB4 is a widely used convolutional architecture, which has shown great accuracy performance for skin cancer classification on datasets such as HAM10000 and ISIC-2019. For the sake of efficiency, its scaling approach balances the network depth, network width and input resolution. It has a good performance without relying on self-supervised learning and attention mechanisms, but the supervision augmentation combination of affine transforms, flips and USDA, and the data augmentation combination with oversampling and weighted loss, especially the minority class recall improvement, are obviously important. Subsequently, for resource-constrained medical diagnostic systems, EfficientNetB4 globally outperforms larger CNNs in terms of model performance and has a relatively low parameter footprint.

SwinV2:

SwinV2 is a self-attention-based hierarchical transformer that can provide powerful feature representation for multiclass skin lesion classification. It uses non-overlapping windows for input images, and calculates the local attention inside the local window, and then slides the window to exchange the global context information, which is quite suitable for variable lesion shapes and sizes. The results show that SwinV2 has good class balance through its application with focal loss or weighted cross-entropy for deep learning, and it usually has a high accuracy.

ConvNeXt:

ConvNeXt is a next-generation convolutional architecture that is inspired by transformer design, integrating convolutional transformation blocks, called ConvNeXt blocks, responsible for powerful local representation extraction with separable self-attention blocks. Furthermore, recent hybrid systems use ConvNeXt as a backbone, hybridized with transformer stages or focal attention, producing state-of-the-art performance in the skin lesion analysis task, and typically yielding better results than pure ViT or traditional CNN baselines in terms of overall accuracy and macro-F1. Various methods using ConvNeXt for skin cancer classification have shown even better interpretability when combined with a spatial attention module or GradCAM visualization, indicating that lesion features account for model predictions.

ResNet101:

ResNet101 is a moderately deep residual network that has been widely used as a baseline architecture for many medical imaging studies such as melanoma detection and HAM10000 classification. Its skip connections and deep feature operations allow, together with the fine-tuning for skin lesion classification tasks, a high accuracy and a stable training, especially for skin lesion classification. Interpretability is often accomplished using external tools like GradCAM, which helps provide a post-hoc visualization of the billions of neurons that have contributed to classification.

DenseNet201:

DenseNet201 is a deep convolutional neural network that has a high density of connections between the convolutional blocks. Generally, each layer receives the feature maps from all previous layers; it is found that this structure could improve the propagation and reuse of features in the visual task. DenseNet201 shows strong generalization and high baseline accuracy with skin cancer datasets and outperforms previous techniques when combined with class balancing methods, including oversampling, even with weighted cross-entropy to address highly balanced lesion data distributions. However, self-supervised training is not included in the standard DenseNet201 pipelines for medical imaging.

EfficientNetB1:

EfficientNetB1 is chosen due to its compound scaling strategy, which achieved high accuracy at low computational cost by balancing depth, width and resolution. Especially for use in clinical applications, where the deployment should be real-time and constrained on resources, it is beneficial to have models capable of extracting hierarchical features from the images efficiently with limited parameters, which is exactly the capability of EfficientNet.

TinyViT:

TinyViT is a compact companion to the Vision Transformer, which means we can expect the tiny model to bring Vision Transformer-level result accuracy in a very space-efficient form, which is especially desirable for mobile/embedded medical imaging systems. Its attention blocks with different scales allow the model to capture complex lesion features with relatively low computational complexity. The strong generalization power of TinyViT can be further improved by its regularization and ensemble training method, which delivers highly consistent performance across different lesion types and image qualities to handle highly imbalanced skin cancer datasets.

MobileNetV3:

MobileNetV3 integrates lightweight architecture innovations, including depth-wise separable convolutions and lightweight attention modules and achieves fast inference and low memory. For this reason, MobileNetV3 is a promising model for point-of-care diagnosis, teledermatology, and smartphone-based screening applications.

ResNet18:

ResNet18 is a traditional convolutional neural network model constructed on the basis of residual learning and is often used as a basic convolutional neural network benchmark in skin cancer classification, transfer learning, ablation experiments, etc. Due to its lightweight nature, it provides fast training convergence and is preferred for low-resource or real-time diagnostic systems.

DenseNet121:

DenseNet121 is a lightweight version of the DenseNet family, adopting the same densely connected block design as DenseNet201, while retaining the efficient gradient flow and feature reuse properties with less depth and parameters. DenseNet121 is the go-to benchmark model because it is known to be reliable at providing strong baseline performance for both binary and multiclass skin lesion classification tasks. Like DenseNet201, DenseNet121 does not come with a self-supervised learning training scheme by default, but it is often fine-tuned from ImageNet weights for medical imaging purposes.

Table 3.1.2.4 contains some benchmark results of deep learning architectures on ImageNet including complexity vs. performance trade-offs. In particular, it compares several models, including ViT-B/16, EfficientNetB4, SwinV2-B, ConvNeXt-B, ResNet101, DenseNet201, EfficientNetB1, TinyViT-5m, MobileNetV3, ResNet18 and DenseNet121 along many metrics like parameters (M), top-1 accuracy (ImageNet), computational cost (FLOPs in G). Our results are used to compare the suitability of each model for skin lesions classification tasks based on efficiency and accuracy.

Table 3.1.2.4: Benchmark result of Deep Learning Architectures on ImageNet

Model	Architecture Type	Parameters (M)	Top-1 Accuracy (ImageNet)	FLOPs (G)
ViT-B/16	Transformer	86.6	82~85%	17.6

EfficientNetB4	CNN	19	82.9%	4.2
SwinV2-B	Transformer	88	84.0%	15.4
ConvNeXt-B	CNN	89	83.8%	15.4
ResNet101	CNN	45	77.4%	7.6
DenseNet201	CNN	20	77.4%	4.4
EfficientNetB1	CNN	7.8	79.2%	0.7
TinyVit-5m	Transformer	5	82.5%	1.3
MobileNetV3	CNN	5.4	75.2%	0.22
ResNet18	CNN	11.7	69.8%	1.8
DenseNet121	CNN	8	74.9%	2.9

3.1.3 Functional and Nonfunctional Requirements

Functional Requirements:

Functional requirements provide basic functionalities, which should be exposed from the system to facilitate a proper classification of skin lesions for meeting this goal. These requirements help to ensure that the system is working correctly and also that it meets the needs of users in a clinical setting.

1. **Data Input and Image Preprocessing:** The system needs to preprocess the dermoscopic images in different formats and consider scaling, normalization, and augmentations on the images. In this procedure, the images are resized to standard dimensions to give uniformity and also to advance the generalization of the model under different image conditions.

2. **Image Augmentation:** The system should include image augmentation techniques like image rotation, image flipping, image contrast and other image transformations to artificially expand the training dataset. This makes the model robust to changes in image orientation and illumination as well.

3. **Training and Fine-tuning:** The system should have the facilities to train with a model in the self-supervised learning (SSL) fashion for feature extraction and fine-tuning to the labeled data. The system needs to categorize seven classes of skin lesions, which means attempting to extract features from dermoscopic images that are useful and optimize these degrees of freedom from labelled data.

4. **Prediction and Classification:** After the system is trained, new dermoscopic images must be classified into one of the skin lesion classes defined by specialists.

The classification must be accurate and reliable, and provide the smallest margin for error for both false positive and false negative errors.

5. **Model Interpretability**: The system should be equipped with model interpretability and allow for understanding the decisions taken by the model. Machine learning models designed to improve radiology reporting, such as Grad-CAM or LIME, should be presented to clarify the system's reasoning behind each prediction and, as a result, establish clarity and trust in the system with radiologists.

6. **Evaluation:** The system should be able to report evaluation metrics such as accuracy, precision, recall, F1-score etc. It also needs to deliver thorough reports about his or her model performance and what they have learned to enable more analysis and improvement.

Non-Functional Requirements:

Non-functional requirements involve quality requirements like performance and security of the system in a real-world healthcare application scenario.

1. **Performance and Speed:** The system should be built to provide a low-latency performance, so it can make predictions quickly and incur minimal delay between image input to the classifier for prediction output. This is important in a clinical setting where the time of the decision for the diagnosis is very critical.

2. **Scalability:** It should be possible to add more and more dermoscopic images to the image generation process without losing performance. It should be able to handle huge amounts of data well in the training and prediction stages.

3. **Accuracy and Reliability:** The system's skin lesion classification accuracy should be high and it should be reliable under various skin conditions and be able to keep the classification system consistent using different datasets and imaging environments.

4. **Security and Privacy:** The system is needed to comply with privacy laws such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) in order to ensure that patient information is kept secure and private. Sensitive information should always be encrypted, stored securely and properly controlled so that unauthorized parties can't gain access.

5. **Maintainability:** The system should be easy to maintain, organize and design. This will ensure an easy way to periodically update, including retraining the model or introducing new features, ensuring that there are no major disturbances on their operations.

6. **Robustness:** The system should be robust enough to handle different qualities of images, as well as lighting and other conditions that may be faced within the real-world of medical imaging. It should have the ability to work with noisy or poor images as well.

7. **Regulatory Requirements:** The system has to satisfy requirements for medical diagnostics and regulatory requirements to ensure that it is suitable for healthcare applications. In keeping with basic legal and ethical requirements for patient safety and the correctness of the diagnosis of the disease.

3.1.4 Data Flow Diagram

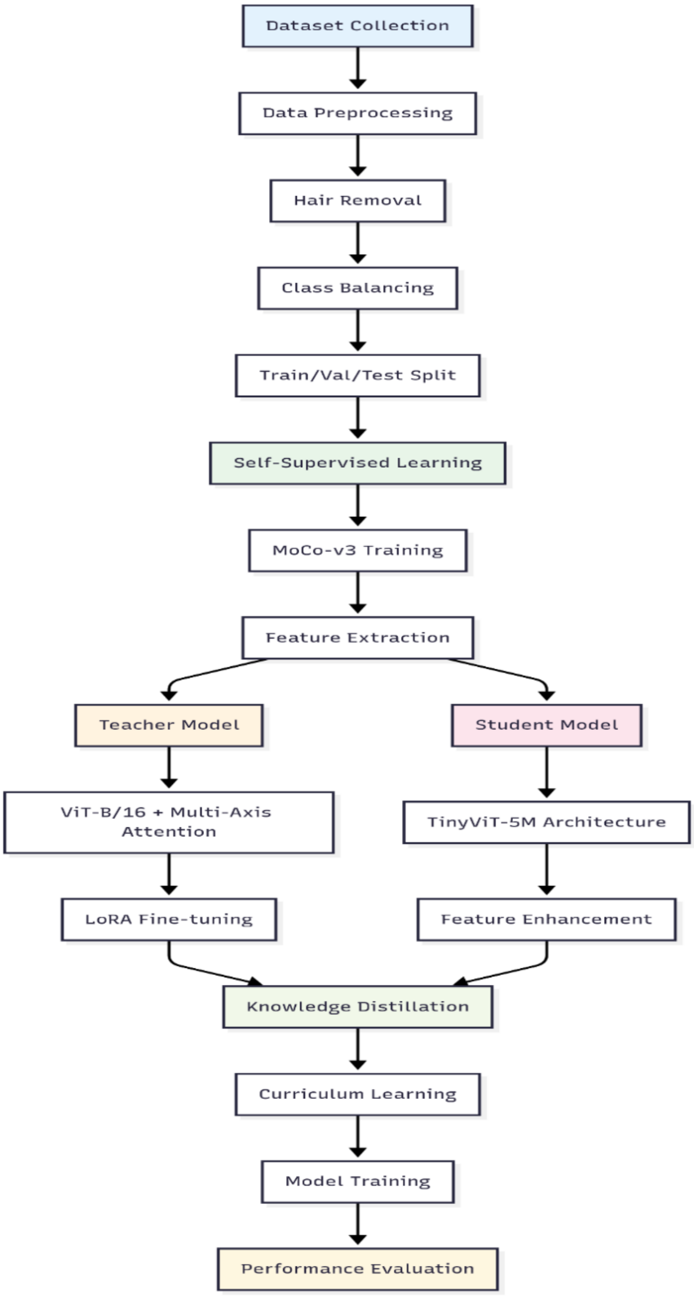


Figure 3.1.4: Data Flow Diagram of Proposed Research Methodology

3.2 Detailed Methodology and Design

In an effort to develop a skin cancer classification system to be more efficient, various approaches were researched before landing on the final method. The following considerations were considered in making the final decision on methodology:

1. **Hair Removal Technique Selection:** In the process of hair removal technique investigation, several methods were tried, such as the use of different razors and methods. The Dull Razor Algorithm was selected not only because it was an efficient algorithm for different hair types, but also because it had a faster processing time and was a batch-processing algorithm suitable for bulk processing. Other methods either hadn't demonstrated consistent performance or took longer to run and thus were less appropriate.

2. **Handling Class Imbalance:** There were several methods that were taken into consideration to handle class imbalance in the dataset, like SMOTE (Synthetic Minority Over-sampling Technique). However, SMOTE resulted in overfitting and went for a Class-weighting system with Focal Loss. This approach essentially balanced the issue of class imbalance without overfitting, because it gives precedence to the classes that are more difficult to classify.

3. **Self-Supervised Learning (SSL) with MoCo-v3:** When it came to self-supervised learning, MoCo-v3 was chosen over SimCLR due to its previous training speed and efficient fine-tuning speed. MoCo v3 uses a contrastive learning approach to pretrain the teacher model using unlabeled dermatoscopic images so that it can learn powerful feature representations focused on dermatology-specific patterns, such as lesion shapes and borders. This choice helped to improve the model's ability to capture critical features.

4. **Teacher Model Selection:** The Vision Transformer (ViT) was used as the teacher model because it has a spatial context model capable of kernels in medical images and also context in medical images, important for skin lesion detection. Another transformer model known as Swin-V2 was tested and was found to underperform compared to ViT because of the differences in architecture and training dynamics of the two models.

5. **Self-Attention Mechanism:** Focal-Self-Attention was also investigated but caused a lot of overfitting due to the importance given to the local features, which makes it hard for generalization. To remedy this, multi-axis attention was added, balancing fine-grained information from local data when focusing on finer scales, and coarser information that captures correlations between pixels for generalization. Low-Rank Adaptation (LoRA) was added to keep the number of learnable parameters down, which not only allows for more efficient training, but also doesn't sacrifice performance.

6. **Student Model Selection:** Following experimentation with different models such as EfficientNet-B1, MobileNetV3 and DenseNet-121, the model TinyViT was selected for the student. TinyViT has been chosen because of its compatibility with the teacher model based on ViT, because the transformative basis of both models makes it easy to perform this knowledge distillation. In addition, TinyViT provides a computational efficiency with a high-performance, which makes it suitable for deployment in restricted-resource environments.

7. **Knowledge Distillation with Curriculum Learning:** Standard Knowledge distillation was tried but it caused it to overfit and have poor generalization, and especially uneven performance for classes. To get around this, Curriculum-Based learning, where the model begins by training on simpler examples before being tested with more complex cases. This ends up as improved learning efficiency, overfitting and generalization results.

8. **Final Methodology:** The final implemented method is a combination of ViT, Multi-Axis Attention, LoRA, MoCo-v3 SSL & TinyViT for the student model. This combination gives a balance to perform the model efficiently and computational cost while covering up issues such as class imbalance, over-fitting, computational cost. The methodology, based on good knowledge transfer, is especially suitable for skin lesions for medical measures of classification.

3.3 Project Plan

This project focuses on the construction of deep learning model for efficient skin cancer classification with attention to accuracy and generalization to diverse data sets. The model will integrate self-supervised pretraining, medical-aware augmentation and knowledge distillation techniques to address important problems such as skewed data and computational constraints. The main activities under the plan are to:

1. **Data collection and preprocessing:** The collection and the preprocessing of the clinical images of skin lesions to include diversity and larger variety of real case scenarios.
2. **Model development:** Design and optimization of deep learning using advanced techniques, such as self-supervised pretraining and medical-aware augmentation, to enhance model accuracy and generalization and knowledge distillation to achieve a more efficient deep learning model for clinical deployment.
3. **Model evaluation:** Testing and validating the model on different data pertaining to the evaluation of the model's accuracy, robustness and generalization capabilities.
4. **Clinical deployment readiness:** Making sure the model is ready for use in healthcare systems, optimizing the model's performance for skin cancer diagnostics in the real world.

3.4 Task Allocation

The Gantt chart in Figure 3.4 indicates the timeframe of the major activities of each period of the research project, from week 12 to week 46.

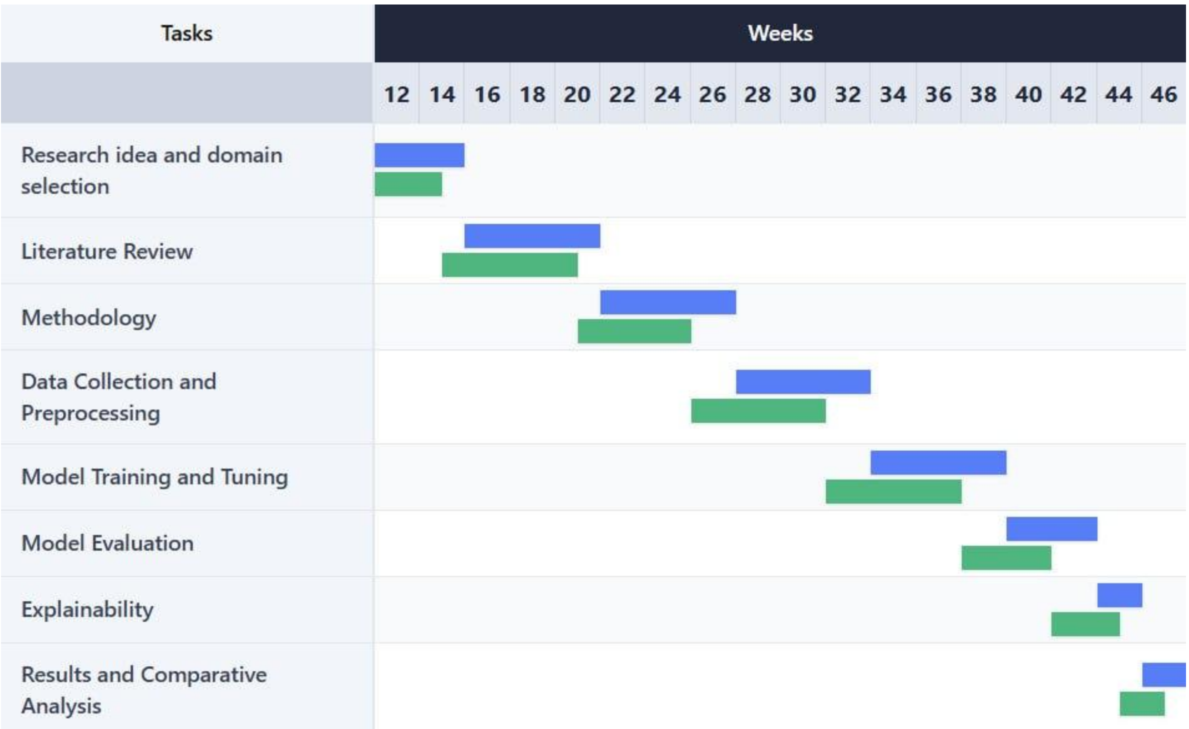


Figure 3.4: Task Allocation

3.5 Summary

This section summarizes a systematic method of classifying skin cancer in a single step, which combines data preparation, self-study, and knowledge distillation. It focuses on utilizing three diverse datasets with Dull Razor Algorithm and dual-level balancing property to solve the class imbalance issue, MoCo-v3 for feature extraction, and ViT-B/16 teacher model with Multi-axis attention and LoRA and it distills to TinyViT-5M student model using Multi-component loss function and curriculum-based learning. The technique provides performance efficiency and reliability that are assessed using some performance metrics such as accuracy, precision, recall, F1, confusion matrices, and ROC-AUC curve metrics, making it accessible to the medical domain.

Chapter 4

Proposed Knowledge Distillation Framework

This chapter proposes a comprehensive skin cancer classification framework based on self-supervised learning and elaborates teacher and student models with an improved knowledge distillation process. Following feature extraction based on MoCo-v3, we further develop and combine these components, namely multi-axis attention and LoRA, optimize teacher and student models, and arrive at an end-to-end efficient classification system, which is a candidate for medical application.

4.1 Introduction

In this study, a simple yet powerful framework to develop an efficient and effective skin lesion classification system is presented. It is built from the perspective of combining self-supervised learning, an advanced teacher and student model, and an enhanced knowledge distillation process. The method begins with Momentum Contrast version 3 (MoCo-v3), which applies momentum-based contrastive learning to learn efficient feature representations from unlabeled dermatological images. These features are exploited to initialize the teacher model, which is a Vision Transformer Base-16 (ViT-B/16) backbone equipped with Multi-Axis-Attention and Low-Rank Adaptation (LoRA). The student model is based on TinyViT-5M architecture, which is optimized to classify seven lesion types with lower computation. To deal with the tradeoff between accuracy and efficiency, the distillation process adopts the multi-component loss function, adaptive weights and curriculum learning, so that the system can be applied to the resource-constrained medical environments.

4.2 Self-Supervised Learning for Feature Extraction

4.2.1 Self-Supervised Learning

Self-supervised learning (SSL) allows the model to learn meaningful representations from unlabeled data, which is particularly helpful when we only have limited amounts of labeled data. In SSL, part of the input is used to predict other parts of the input, so the model can learn to capture important features of the input, including shapes, textures, and boundaries. These learned features are also important for applications such as image classification. By learning strong representations, SSL helps the model to generalize the downstream tasks with a little amount of labeled data. In this work, we use MoCo-v3 for SSL with contrastive learning to pretrain a ViT-B/16. The model compares augmented versions of the same image as positive pairs and discriminates them from negative samples, allowing the model to construct strong feature representations from unlabeled dermoscopic images. Figure 4.2.1 shows the Self-Supervised Learning Process Using MoCo-v3 for Domain-Adaptive Feature Extraction.

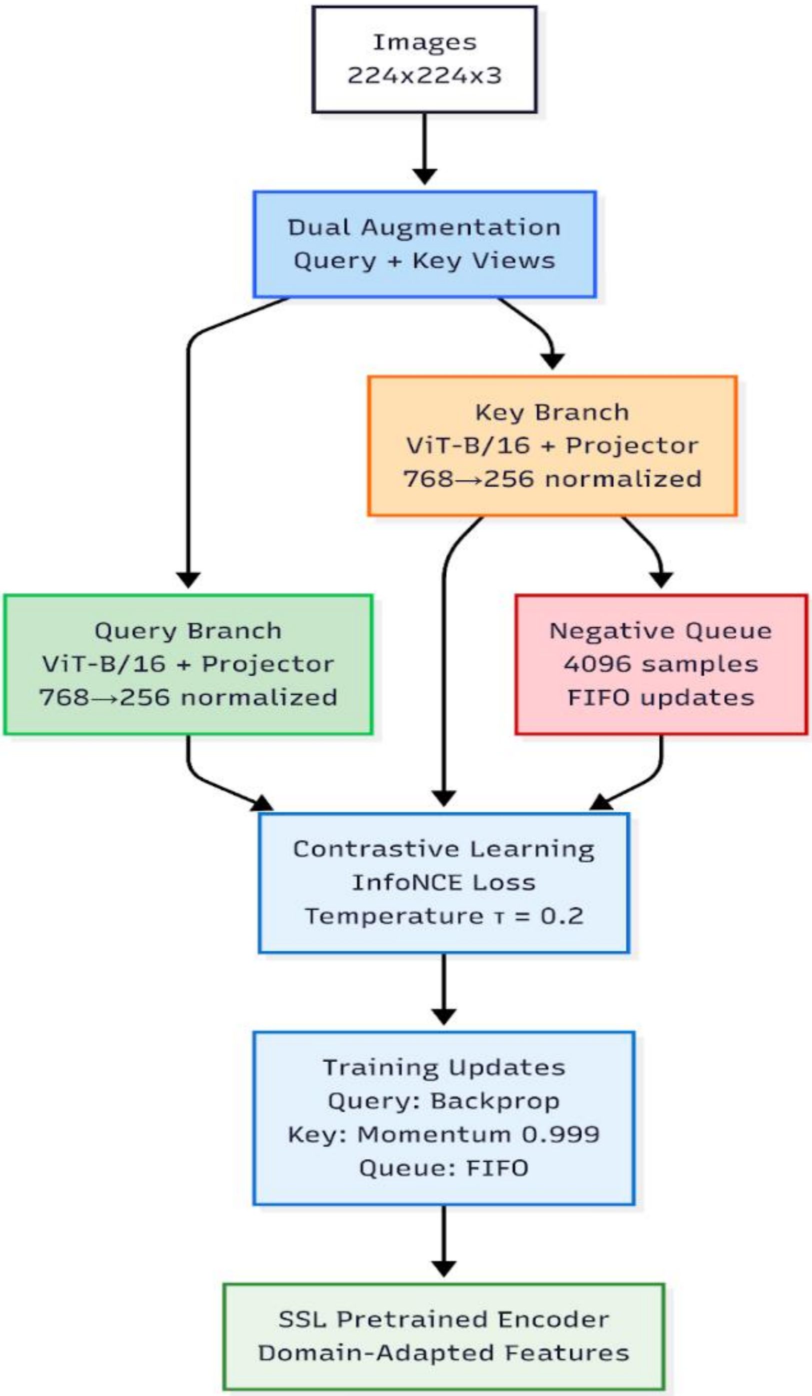


Figure 4.2.1: Self-Supervised Learning Process Using MoCo-v3

4.2.2 MoCo-v3 Architecture and Methodology

The MoCo-v3 architecture uses two identical ViT-B/16 encoders, called the query encoder and the key encoder. The current batch is passed through the query encoder to obtain query embeddings and also through the key encoder to obtain key embeddings that are used as positive or negative sample embeddings in contrastive learning. The momentum update mechanism keeps parameters of the query encoder close to parameters of the key encoder, which is conducive to robust and stable feature representation. Both encoders are trained on ImageNet without classification heads, which results in feature vectors of 768 dimensions. These vectors are fed to two-layer MLP projection heads, consisting of linear layers, batch normalization, ReLU activation, and a final linear layer, which produces 256-dimensional normalized embeddings that are used for contrastive learning. The softmax distribution is scaled in the loss function by a temperature parameter $\tau = 0.2$. The key encoder is updated continuously by the momentum mechanism, so that no sudden changes occur:

$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q \tag{2}$$

where θ_k and θ_q are the parameters of the key and query encoders, respectively, m is the momentum coefficient (0~1) for the temporal consistency.

In the forward pass, two augmented views for each image are generated, one is query view and another one is key view. The query view is passed through the query encoder to generate a 768-dimensional feature vector, which is further projected to a 256-dimensional embedding q by the MLP. The key view is encoded by the key encoder (no gradient updates) into a 256-dimensional embedding k. The positive logit is the dot product similarity between q and k:

$$l_{\text{pos}} = \sum_{i=1}^{N} q_i \cdot k_i \tag{3}$$

where N is the batch size, q_i and k_i are the query and key embeddings for the i-th sample. If the query embedding is compared to a queue of 4,096 previous keys, it creates a matrix of similarities and the logits are negative. The summed logits are scaled by τ:

$$logits = [l_{\text{pos}}, l_{\text{neg}}]/\tau \tag{4}$$

This scaling tends to sharpen the distribution toward differences between positive and negative pairs. The cross-entropy based contrastive loss, which is computed as follows, maximizes similarity between the query and positive keys and minimizes dissimilarity with negative keys:

$$L = -\log\left(\frac{\exp(\text{sim}(q,k^+)/\tau)}{\exp(\text{sim}(q,k^+)/\tau) + \sum_{k^-} \exp(\text{sim}(q,k^-)/\tau)}\right) \tag{5}$$

where k^+ is the positive key encoding and k^- is the negative key encoding.

4.2.3 Self-Supervised Learning Integration in Teacher Model

To add the self-supervised learning on the teacher side, the weights of the teacher backbone are initialized with those of the MoCo-v3 encoder pretrained on dermatological images. This gives domain-specific feature representations, which outperform standard ImageNet pretraining for the medical domain. The transfer checks for dimensional compatibility between source and target models and loads only matching weights:

$$W_{\text{teacher}}^{(i)} = F\left(W_{\text{SSL}}^{(i)}\right), \forall i \in T \tag{6}$$

where $W_{\text{teacher}}^{(i)}$ represents weights in the teacher's i-th layer, $W_{\text{SSL}}^{(i)}$ represents the weights from the MoCo-v3 encoder, and F only passes weights that are compatible. This makes the teacher dermatologically informed, which improves classification performance.

4.3 Teacher Model with Multi-Axis Attention and LoRA

4.3.1 Teacher Model Architecture

The teacher model is based on a ViT-B/16 backbone, which is suitable for medical image classification as the ViT backbone learns local features, as well as global relationships using self-attention. It can take 224x224x3 images and partition them into 16x16 patches, while preserving the spatial relationships and modeling dependencies across the image. The images are converted into embedded patch streams, which maintain important diagnostic information for the task, using a transformer-based image encoder. The ViT-B/16 backbone effectively learns multi-layered representations of features required for capturing complicated patterns from medical images. Its attention mechanism provides an attention to important image regions offering both local and global context, which is essential for tasks such as skin lesion classification. The proposed design is able to extract spatial and contextual information for the model, enabling accurate medical image analysis.

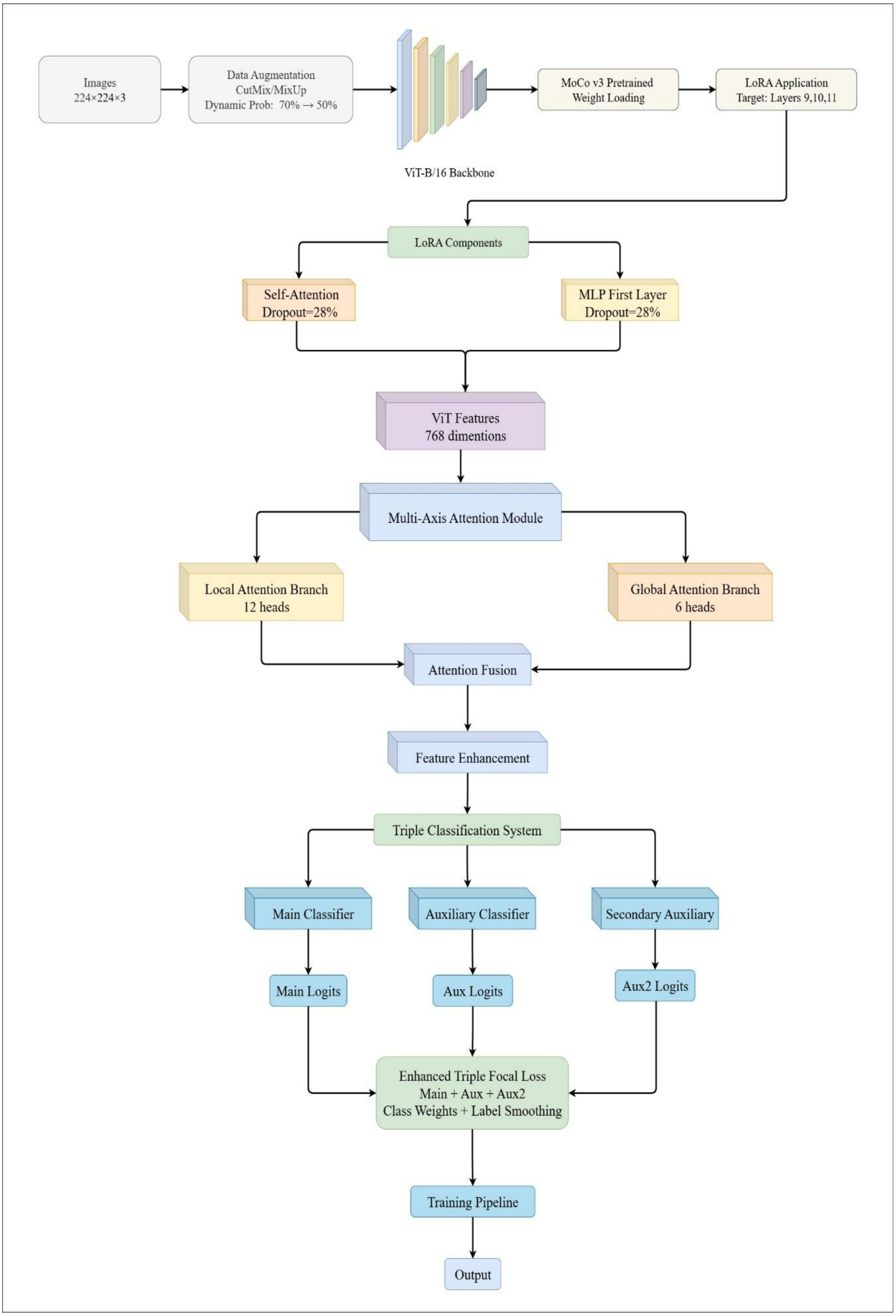


Figure 4.3.1: Teacher model architecture with Multi-Axis-Attention and LoRA

4.3.2 Low-Rank-Adaption Method

To make the fine-tuning process more efficient, LoRA [37] is integrated into the teacher model, which means only a subset of weights is updated instead of updating all weights. The method particularly focuses on the specified layers, such as self-attention and multilayer perceptron (MLP) layers, by incorporating low-rank matrices A and B for performing task-specific adaptation. Due to this, only these matrices are updated during training, and the pretrained ViT weights remain the same. In the teacher model, LoRA works in the following way:

$$W = W_0 + \alpha \cdot A \cdot B \tag{7}$$

where A and B are the low-rank matrices as proposed in LoRA, and W_0 is a pretrained fixed weight matrix, and α is the scaling factor for controlling the strength of the adaptation.

This replaces the projection matrix used in the self-attention layer and the first linear transformation in the MLP layer of the backbone for ViT. It provides the opportunity for the model to learning new task-specific features without changing most original parameters very much. To shrink the updates to these matrices and thus reduce the number of parameters we train, regularization can be added, thereby increasing the training efficiency. After applying LoRA, the pretrained weights of the teacher model come from the MoCo-v3 self-supervised encoder, and the low-rank matrices are trained to adapt to the target task. This hybridity enables the teacher to expand upon previously learned knowledge as well as allow for flexibility in response to new challenges without the need for the entire model to be retrained.

4.3.3 Multi-Axis-Attention Mechanism

The teacher model consists of a Multi-Axis-Attention [38] mechanism to capture both local and global patterns in the medical image data. This mechanism is based on two image feature processing channels, so that the model is able to pay attention to detail at the local level and the large-scale level in order to integrate the local and global patterns in the medical image data. A multi-axis attention mechanism is added to the teacher model. This mechanism can capture image features via two streams, which allows the model to pay attention to fine details locally and the bigger picture globally as illustrated in Figure 4.3.3.

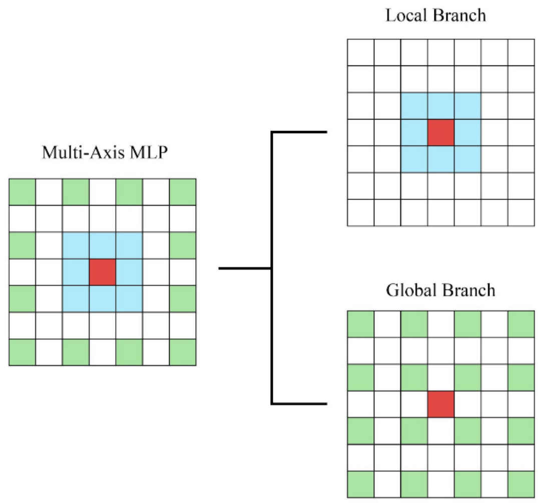


Figure 4.3.3: Multi-Axis-Attention Mechanism

Local Attention Equation:

$$y_{\text{local}} = \text{MultiheadAttention}(x, x, x) \tag{8}$$

Where:
x (Query): This is the input on which the attention mechanism pays attention. In self-attention, the query is equal to the key and value representation from the previous feature sequence of layers.
x (Key): This too is extracted from the input, compared to the query and used to calculate attention scores, which are the weight of each element.
x (Value): Also sampled from the input features, the value is multiplied by the attention scores from the check between the query and the key, then their sum is the output of the attention layer.

Global Attention Equation:

$$y_{\text{global}} = MultiheadAttention(y_{\text{local}}, y_{\text{local}}, y_{\text{local}}) \tag{9}$$

Where:
y_{local} (Query, Key, Value): We put the output of the local attention as query, key, and value, respectively, in global attention. The goal is to capture richer contextual information, in addition to the already processed feature representations, thus allowing the model to attend to long-range dependencies inside the image.

Final Output Equation for Multi-Axis-Attention:

$$y = LayerNorm(y_{\text{global}} + y_{\text{local}}) \tag{10}$$

Where:
y_{global} : The result of the global attention path
y_{local} : The output of the local attention path
$LayerNorm$: Layer normalization is used to normalize the summed results from local and global attention to stabilize training, and further to facilitate convergence.

Multi-axis attention uses two basic paths: local attention and global attention. Local attention allows us to attend to the fine spatial details by aggregating small part-level patterns captured through a self-attention function. Global attention, on the other hand, uses the output of local attention as query, key, and value to learn some higher-level, more global pattern for the image. By bringing these two attentions together, the model learns the fine-grained structure along with the global context, as needed for tasks such as lesion boundary recognition and global contextual understanding of an image. The generated feature representations further improve the effectiveness of the model for the classification of medical images.

4.3.4 Regularization

In this model regularization is implemented using L2 regularization, which helps to prevent overfitting with good generalization. At first, strong regularization is introduced to prevent overfitting problem, and the stronger the accuracy of the validation data, the smaller the scale of the regularization. This way the model is able to do fine-tuning of its parameters effectively and it also remains well-generalized to make a prediction on some unseen data. mIn addition, LoRA (Low-Rank Adaptation) serves as a natural regularization method. By instead introducing low-rank adaptations to these pre-trained layers, LoRA reduces the quantity of trainable parameters, without impairing the model's performance. In doing so, this method leads to a more efficient model without model-overfitting or excessive model complexity, ensuring the model picks up the essential characteristics of the dataset. Furthermore, it allows us to train quicker and to alleviate some of the computation requirements.

4.4 Student Model with Knowledge Distillation

4.4.1 Student Model Architecture

The student model is based on TinyViT-5M backbone, as shown in Figure 4.4.1.1. It is designed in a four-stage structure with channel sizes from 64 to 320 that allows for multi-scale feature extraction, which is the key for medical image analysis. It replaces global self-attention with local attention as the main mechanism for computation efficiency, and employs convolution with a mobile-friendly architecture to reduce parameter usage. The model has basic building blocks of the patch embedding modules, TinyViT blocks, MBConv blocks for mobile purposes, and the patch merging layers to achieve smooth transitions between stages. While the texture of these elements is designed to capture fine-grained local information such as the boundary between lesions, they are also designed to retain broader contextual information relevant to dermatological assessments. The feature processing pipeline uses two stages, with layer normalization and adaptive dropout regularization to increase the output dimensions from 320 to 512.

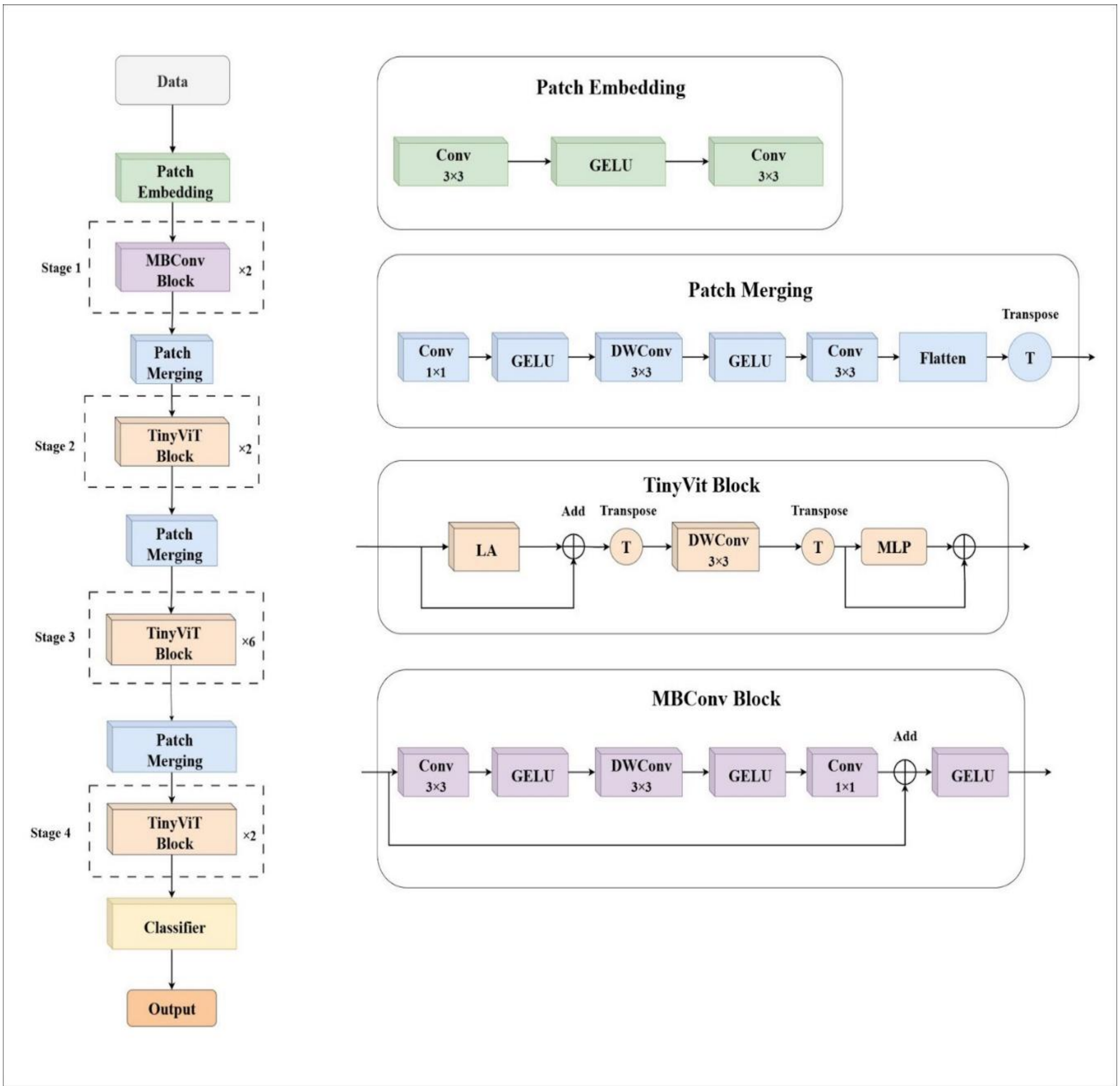


Figure 4.4.1.1: TinyViT Backbone Architecture

Multi-head classification helps to increase training stability and performance. This consists of a main classification branch, a secondary branch for gradient flow augmentation, and a confidence estimation head for uncertainty, which is very important for medical applications. The architecture is efficient while enhancing the resilience of TinyViT-5M for complex clinical image classification tasks. Custom architecture is shown in Figure 4.4.1.2.

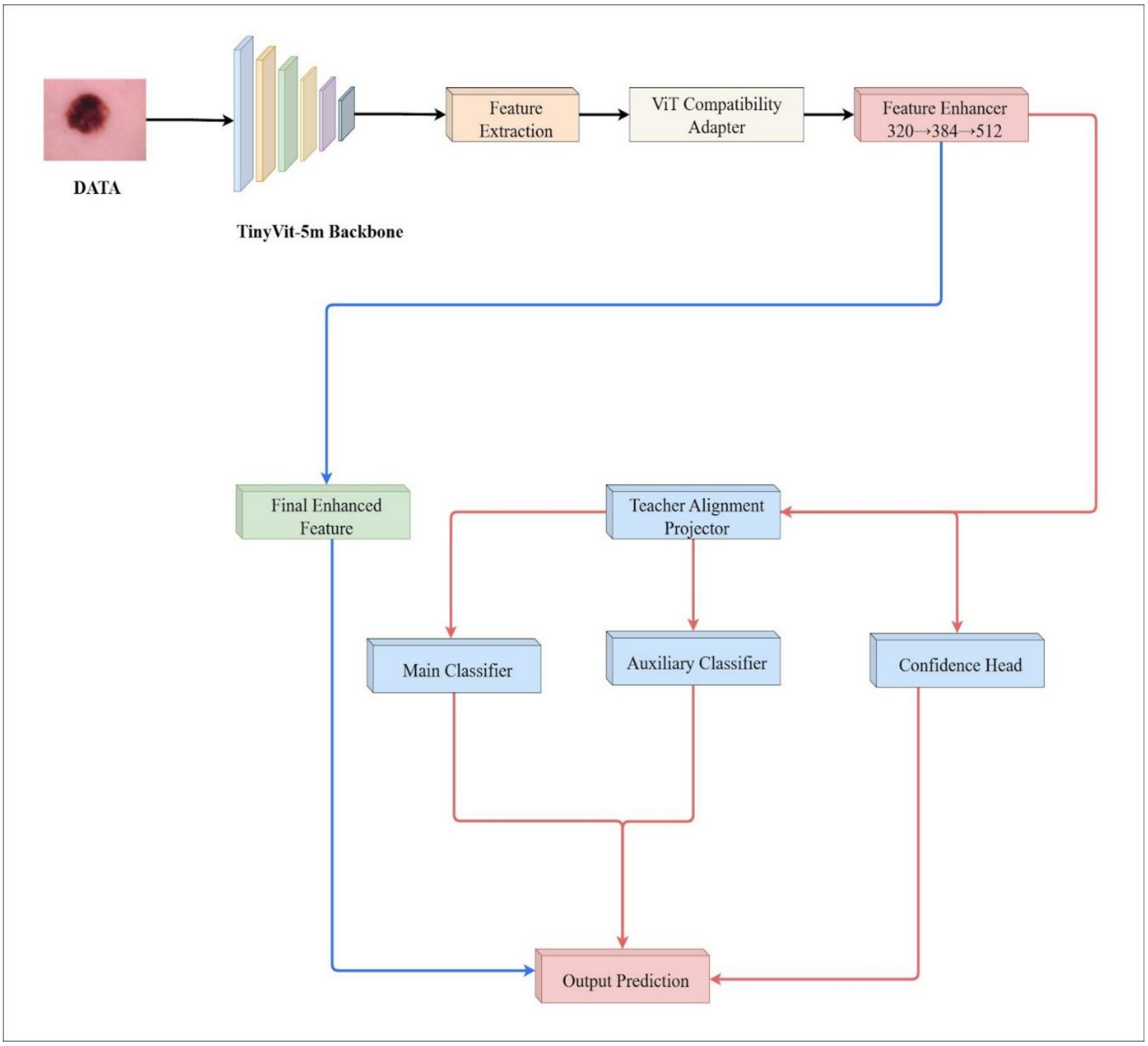


Figure 4.4.1.2: Student Model Custom Architecture

4.4.2 Knowledge Distillation Methodology

Knowledge distillation [39] allows a much smaller student model to learn from a much larger and pretrained teacher model. The teacher is trained on large amounts of data and mentors the student by passing along its predictions and internal representations developing within its features. This enables the student to reproduce the teacher's knowledge using fewer parameters and lower computational demands. As illustrated in Figure 4.4.2, the knowledge transfer is achieved by means of multiple loss functions, adaptive learning-weighting, and curriculum learning strategies, which also allow for efficient learning from the teacher even when relying on the computational savings afforded by a smaller model. The student learns not only the teacher's conclusions, but the patterns of reasoning employed by the teacher. Due to its advantages of learning effectiveness and computational efficiency, this is more valuable for medical image classification, where both accuracy and deployment efficiency are of great importance.

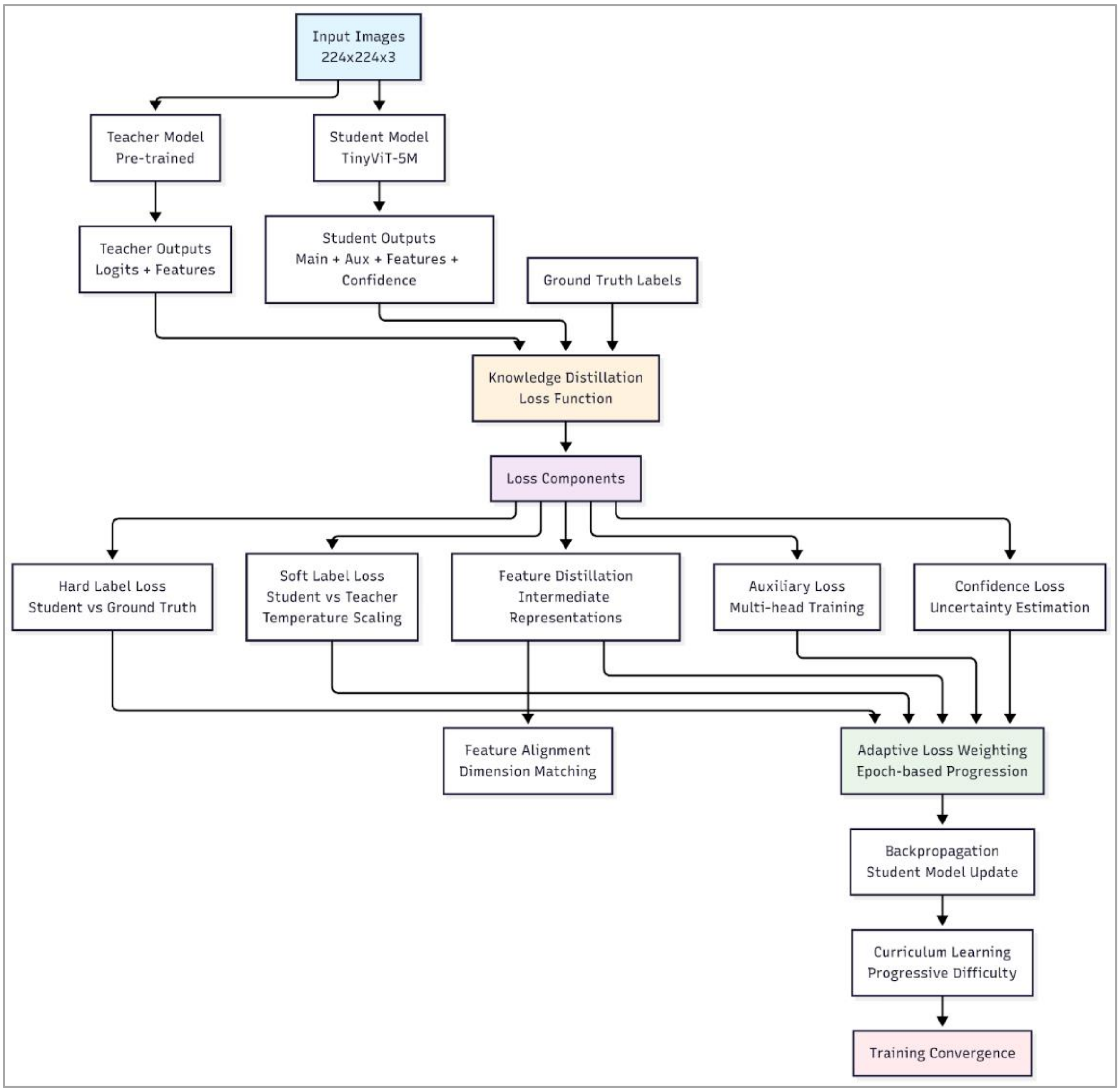


Figure 4.4.2: Knowledge Distillation Architecture Flow

4.4.3 Knowledge Distillation Loss Function

Knowledge Distillation Loss function is designed to guide the student model by assembling a variety of components together so it can learn well from both the real labels and the teacher model. The first part, Hard Label Loss, applies cross-entropy to make sure the student predicts the correct class labels based on the ground truth of the dataset. This is for better accuracy of classification and will help the student to discover the true data. The second part is Soft Label Loss, where it attempts to make the student follow the teacher's softened predictions. In particular, the Kullback-Leibler divergence (KL) between student and teacher outputs is computed between the teacher's outputs masked with the temperature scaling factor (T) and the student output. This helps the student to not only know the expected classes, but to reinforce the teacher's adherence logically and emotionally to the rationale of the decision made. The third part aims for the student to learn the teacher's internal feature representations through the Feature-

Level Distillation Loss that minimizes the Mean Squared Error (MSE) between the representations. This leaves the student with a drive to notice more complex hierarchical patterns present in the data and to understand it at a higher level. The final ingredient, Confidence Regularization, uses the binary cross-entropy to compare the student's estimated confidence against the truth and makes sure the model's confidence is in line with the truth. These components, together, allow the student to learn from the teacher's predictions and internal findings and thus enhance the performance and generalization capabilities. The total Knowledge Distillation Loss can be defined:

$$L_{KD} = \alpha L_{hard} + \beta L_{soft} + \gamma L_{feature} + \delta L_{confidence} \tag{11}$$

Where α, β, γ, and δ are weighting factors that are updated during training.

4.4.4 Curriculum Based Progressive Distillation

Curriculum learning presents training samples in a step-by-step fashion, from the easy to the more difficult. At first, the student model uses simple examples for which the teacher is very sure of its predictions. As training progresses, more and more difficult examples are introduced. This type of preparation develops a patience to prevent overfitting in the beginning and allows the student to adapt gradually for difficult cases. The degree of difficulty of each example is set according to the teacher's confidence, with a high confidence level corresponding to easily solved problems, and the lower the confidence, the more difficult the problems. It enables efficient memory training where the teacher's confidence is used to dynamically select training examples, eliminating the need to copy the data for various levels of difficulty.

4.4.5 Knowledge Distillation Flow in Teacher-Student Model

In knowledge distillation, the teacher model is used to generate soft predictions and feature representations to guide the student towards the correct predictions, using its weights frozen. The teacher transfer its knowledge and internal features to the student, so that the student can benefit from it without having to retrain too much. The student's goal is then to minimize the variance between its own predictions and features, and those of the teacher, replicating the performance of the teacher while retaining the benefits of a more compact design. The student model is trained with both the ground truth hard labels and the teacher soft labels. It begins by driving learning of the correct class predictions from hard label loss. Over time, the focus moves to the teacher's soft predictions as well as the teacher's implied internal characteristics. Curriculum learning starts from the simpler cases and proceeds to contact increasingly harder cases, which guarantees that the student learns enough without overfitting. This systematic approach enables the student to receive the teacher's knowledge in an orderly and effective manner.

4.5 Summary

In this chapter, we present a framework for skin lesion classification that integrates self-supervised learning, knowledge distillation and efficient model design. A teacher-student model based on a multi-axis attention and LoRA enhanced ViT-B/16 backbone is used as the strong feature extractor, where the MoCo-v3 is used to extract the strong features from the unlabeled dermatology images to be used as an input to the teacher model. The student model is trained using curriculum learning strategy based on TinyViT-5M considering both classification accuracy while maintaining computational efficiency. Furthermore, the knowledge distillation process relies on an adaptive weighted multi-part loss function to achieve effective knowledge transfer while maintaining low resource consumption, which makes the system suitable for operation in medical environments with limited computation power.

Chapter 5

Implementation and Results

This chapter presents the implementation and results of the skin cancer detection system, which includes the environment setup, model training process, performance evaluation metrics, system comparative analysis and results discussion. It provides a detailed summary of the experimental setup, the training process, the evaluation results and a summary to highlight the effectiveness of the system for improved accuracy of detection.

5.1 Environment Setup

5.1.1 Hardware Setup

The model was trained on a system that had an Nvidia GeForce RTX 3060 GPU with 12 GB of VRAM and an Intel Core i7-12700, an 8-core processor, with 16 GB of RAM. This configuration ensured dealing with the computational requirements of the machine to process the training of the model, especially with large data sets of medical images.

5.1.2 Software Setup

Python 3.8 was used as a primary programming language due to its flexibility, a large number of library support. PyTorch and TorchVision were used for model development, training and evaluation. PyTorch provided the versatility in building complex architectures and TorchVision provided tools to handle image reading and processing in a simple way. Bands in pandas, numpy processed the big data and CSV files, making manipulation of data easy. Pillow-supported Image preprocessing tasks, which can easily be resized, converted, and augmented to enhance the generalization of the model. Scikit-learn helped to split the dataset between train data, validation data, and test data, to ensure that the performance can be assessed robustly. Albumentations was chosen for image augmentation because of its extensive list of transformations designed to increase the diversity and quality of input data during training. The model architectures were then loaded into the Timm library as well as TorchVision to ensure efficient designs in classification tasks. Matplotlib to provide plots to share and track training progress, and TQDM for the real-time progress bar while acoustic models run. For improved interpretability, Grad-CAM were added for insights into the model's decisions. The setup was based on a Windows operating system with the configuration for seamless integration to Python tools. The environment was managed by Conda, and an interactive development and testing environment was available via Jupyter Notebook. CUDA 11.2 contributed to the GPU acceleration provided by the smooth computations effect in Nvidia GPU and on the other hand, PyTorch 1.10.0 is completely compatible.

5.2 Model Training

We carefully conducted our training using a set of trained parameters, as well as a process, in order to achieve promising results to perform and generalize our models in the best possible way to detect skin cancer.

Training Parameters:

- **Data Augmentation:** A Custom pipeline of augmentation, including flipping, rotation, brightness and noise augmentation, was added to the set of training images. The parameters set included as much as 30 degrees of rotation, flipping horizontally or vertically with a probability of 0.5, as well as similarities of the brightness factor ranging from 0.8 to 1.2, and similarities of the Gaussian noise of 0.01 standard deviation.
- **Epochs:** Processing involves 400 epochs of SSL with MoCo v3 training process, 70 epochs for training the teacher model, 60 epochs for implementing knowledge distillation.
- **Batch Size:** A batch size of 64 has been used as a trade-off between computation efficiency and generalization of the model.
- **Image Size:** Input images were scaled down to size 224x224 to try and include as much detail in the image and still keep the computational changes reasonable.
- **Learning Rate:** Learning rate was fixed at $2.5e^{-5}$ to have constant weight adjustment and slower convergence.
- **Weight Decay:** Weight Decay 0.01 was used as a weight decay for the regularization property to avoid the situation of over-fitting that might occur.
- **Regularization Techniques:** To prevent overfitting, a dynamic regularization technique was employed on the basis of an initial probability of 0.7; techniques like CutMix, MixUp were employed. The probability was slowly scaled down to achieve good validation accuracy, and the concentrations are expected to focus on fine-tuning at a later model process training.

Training Process:

- **Data Preprocessing**: Images were resized and augmented using flips, rotations, brightness adjustments and noise to ensure an increase in generalization being exposed to various conditions.
- **Model Configuration:** Model architectures were set up for all initial purposes using parameters to be used for optimal performance. Weights were initialized with normal ones, which are suitable for ReLU layers, and enable stable and efficient training.
- **Training Loop**: The model processed a set of images at a time for each epoch. After

each batch, the loss was calculated and weights were updated via backpropagation in order to improve the performance of the model shortly after each other.

- **Regularization Techniques:** Dropout and L2 Regularization have been utilized in order to overcome the problem of overfitting. Dropout randomly deactivated the units, and L2 regularization reduced the large weights geometry, forcing the more generalized models. Additionally, CutMix & MixUp were used to augment training data by combining multiple images to obtain better variations in the training body of data, therefore reducing over-training.
- **Learning Rate Scheduling:** The learning rate was altered during the process of training. If the validation loss plateaus, the learning rate is reduced so that finer adjustments can be made to optimize the converging process.
- **Curriculum Integration:** A Curriculum approach in which to organize images in their level of difficulty, and move the model in the model of memorizing in a difficult order, learning well.
- **Final Assessment:** Training came to a close with a comprehensive assessment to make sure the model is ready for fieldwork.

5.3 Performance Evaluation Metrics

In this study, a set of performance measures are used to assess the potential of the proposed model and comparative models to classify skin cancer. The metrics allow accuracy, reliability, and clinical relevance analysis, which are necessary for the classification of skin lesions.

Accuracy: This is a measure of the overall correctness of the model prediction. It's the fraction of correctly classified images over the number of tested images, and is an aggregate measure of performance on all classes.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Samples} \tag{12}$$

Precision: Also known as positive predictive value, precision is the ratio of true positive predictions divided by all positive predictions. In the medical imaging industry, precision helps to significantly reduce false positive results that can unnecessarily cause people to undergo unnecessary procedures or worry about their health.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{13}$$

Recall: It is also called sensitivity, and it represents the capability of the model to identify all relevant instances of a class. This is critical in the detection of skin cancer, as false negatives could delay life-saving treatment for melanoma.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (14)$$

F1 - Score: A simple harmonic mean of Precision and Recall, which weights the two metrics equally and is very useful for imbalanced datasets as it represents the performance on minority classes.

$$F1 - Score = 2 \times \frac{Precision \times Recal}{Precision + Recal} \quad (15)$$

Confusion Matrix: Reports a visual summary of the model's performance with respect to the type of classification the model is doing, including True positives, True negatives, False positives, and False negatives for each type of lesion. It helps to discover certain strengths and suboptimal areas, for example, confusion in the identification of similar classes like melanocytic nevi and melanoma.

ROC Curve and AUC: The Receiver Operating Characteristic curve is a plot of the false positive rate on the x-axis and true positive rate on the y-axis using different thresholds being applied. On the other hand, Area Under the Curve (AUC) measures the ability of the model to distinguish between classes; in other words, it measures the discriminatory power of the model. For a multi-class problem, these are computed in a one vs rest way. High values of AUC present good separability and thus require achieving a reliable diagnosis of skin lesions in the clinical environment.

5.4 Comparative Analysis

5.4.1 Experimental Results of Pretrained Model

The experiment was performed using 3 different HAM10000, ISIC2019, and PAD-UFES-20 datasets, which contain various images of skin lesions that belong to several different categories. To enhance the model robustness and decrease the overfitting, the training images were subjected to the data augmentation techniques.

The models used in this first experiment are a set of state-of-the-art pre-trained architectures like EfficientNetB4, SwinV2-b, ConvNeXt-B, ResNet101, DenseNet201, VIT, EfficientNetB1, TinyViT, MobileNetV3, ResNet18 and DenseNet121. These models were evaluated on the basis of their performance in accurately classifying skin lesions in these datasets.

The models were tested with an accuracy being one of the primary metrics and the results were shown in figures of comparison. Figures 5.4.1.1-5.4.1.3 show the accuracy performance of the models for the datasets, giving an insight into which are more efficient in classifying skin lesions. The highest accuracy on the HAM10000 dataset was 87.43% on ViT and the lowest was 74.45% on ResNet18. Similarly, on ISIC-2019, the highest accuracy also came from the ViT model (86.90%) and the lowest was from the

ResNet18 (73.24%). Finally, on PAD-UFES-20, the highest accuracy was 87.87% on the ViT model and the lowest was on the ResNet18, 74.90%.

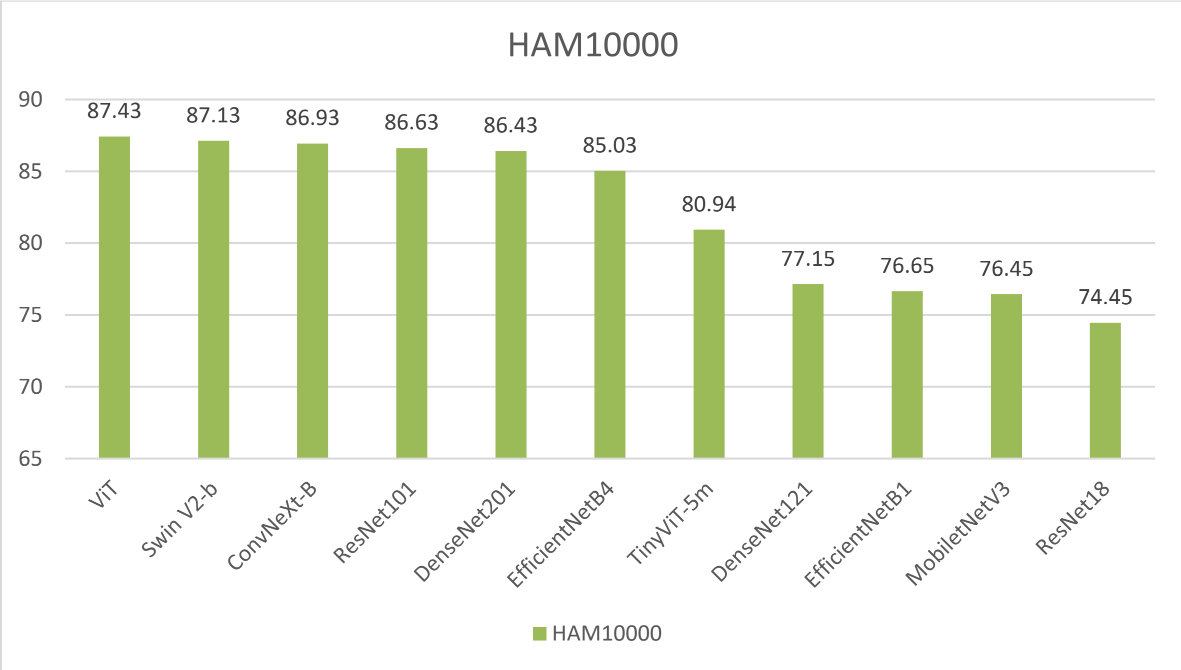


Figure 5.4.1.1: Pre-trained Model Accuracy for HAM10000 Dataset

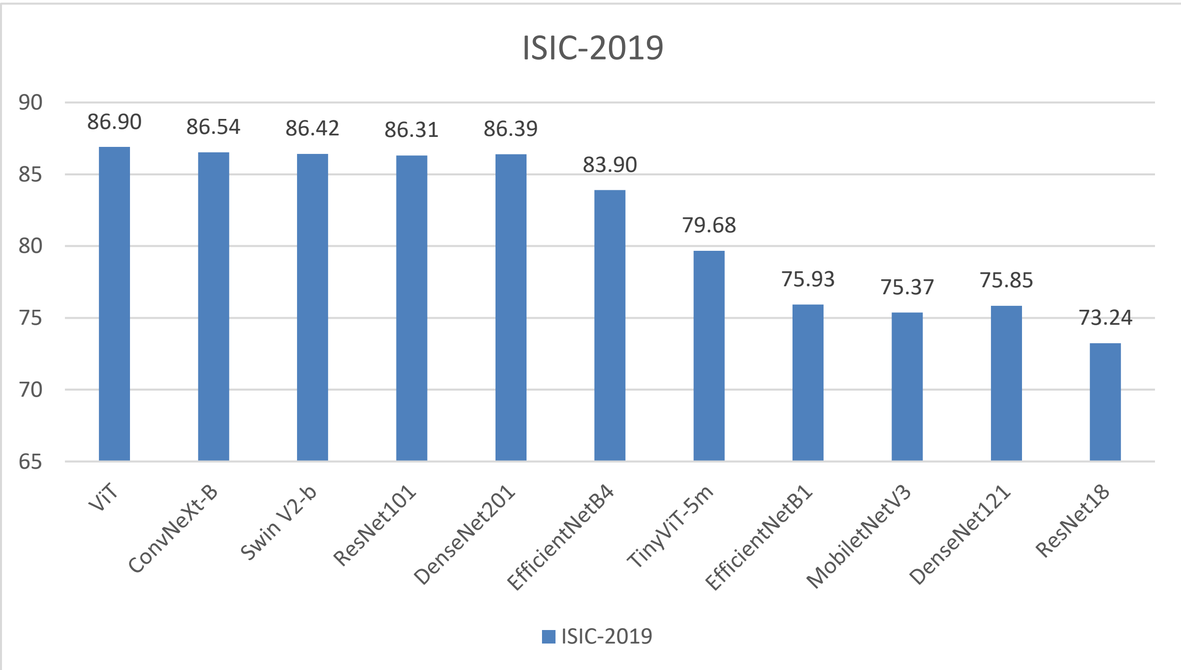


Figure 5.4.1.2: Pre-trained Model Accuracy for ISIC-2019 Dataset

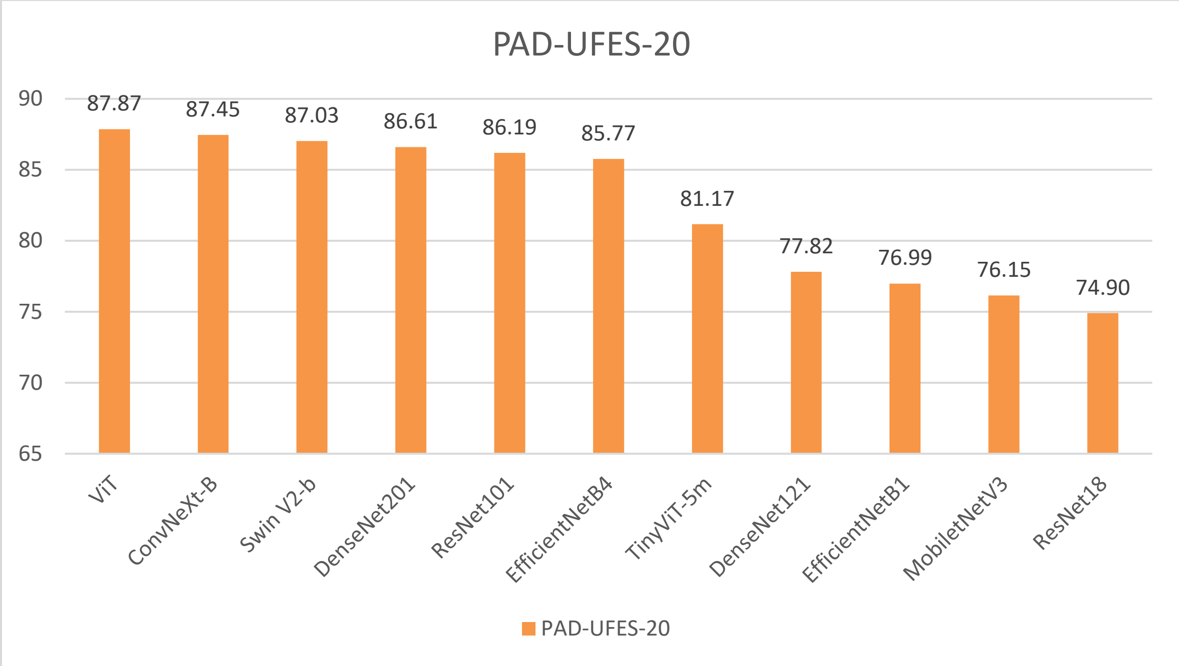


Figure 5.4.1.3: Pre-trained Model Accuracy for PAD-UFES-20 Dataset

5.4.2 Experimental Results of Ensemble Model

In this study, an ensemble model that contains TinyViT-5M, EfficientNetB1, and MobileNetV3 with progressive fusion strength from 0.01 to 0.70 is used. The model was tested on the same three datasets, HAM10000, ISIC-2019 and PAD-UFES-20. On the HAM10000 dataset, the maximum accuracy reached was 84.93%. The accuracy of the ISIC-2019 dataset was 83.31%, whereas the dataset with the highest accuracy is PAD-UFES-20, which had 85.77%. The loss curves for both the training and validation datasets indicate that there is some variance, meaning some overfitting, but also the accuracy on the validation data set was more steady at just over 80%. This means that this model was robust and flexible, which means a pretty much good performance with all three datasets. However, the ensemble model was not that efficient. The three combined models have a total of 16 million parameters and the training process for the three combined models was longer than that of a single pre-trained model.

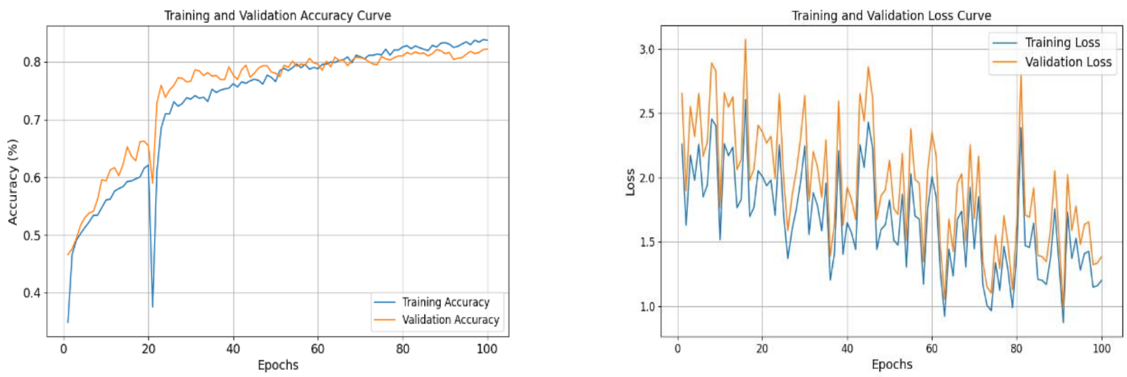


Figure 5.4.2.1: Ensemble Model Accuracy and Loss Curve of HAM10000

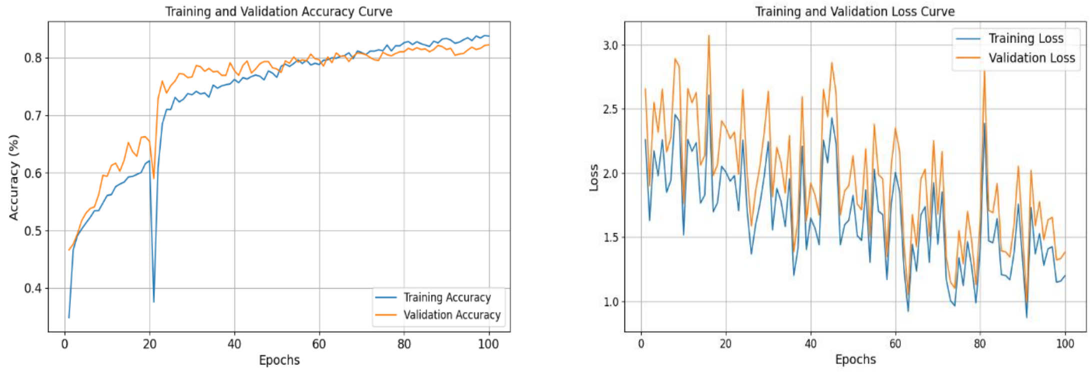


Figure 5.4.2.2: Ensemble Model Accuracy and Loss Curve of ISIC-2019

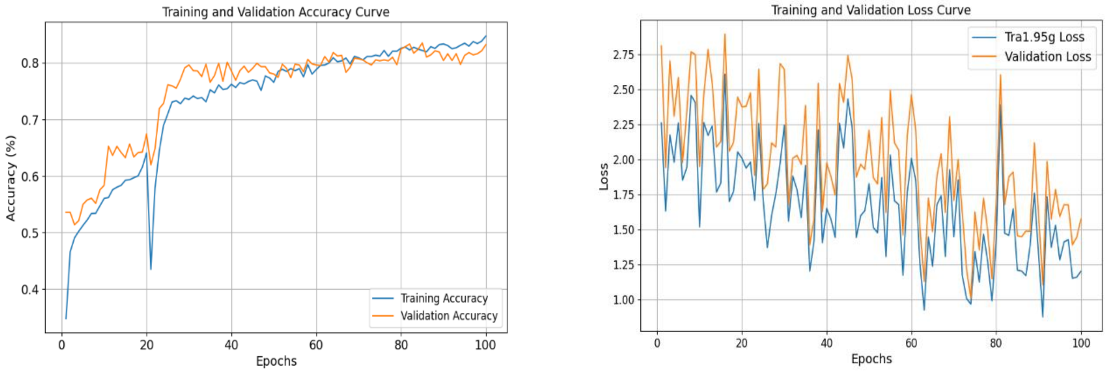


Figure 5.4.2.3: Ensemble Model Accuracy and Loss Curve of PAD-UFES-20

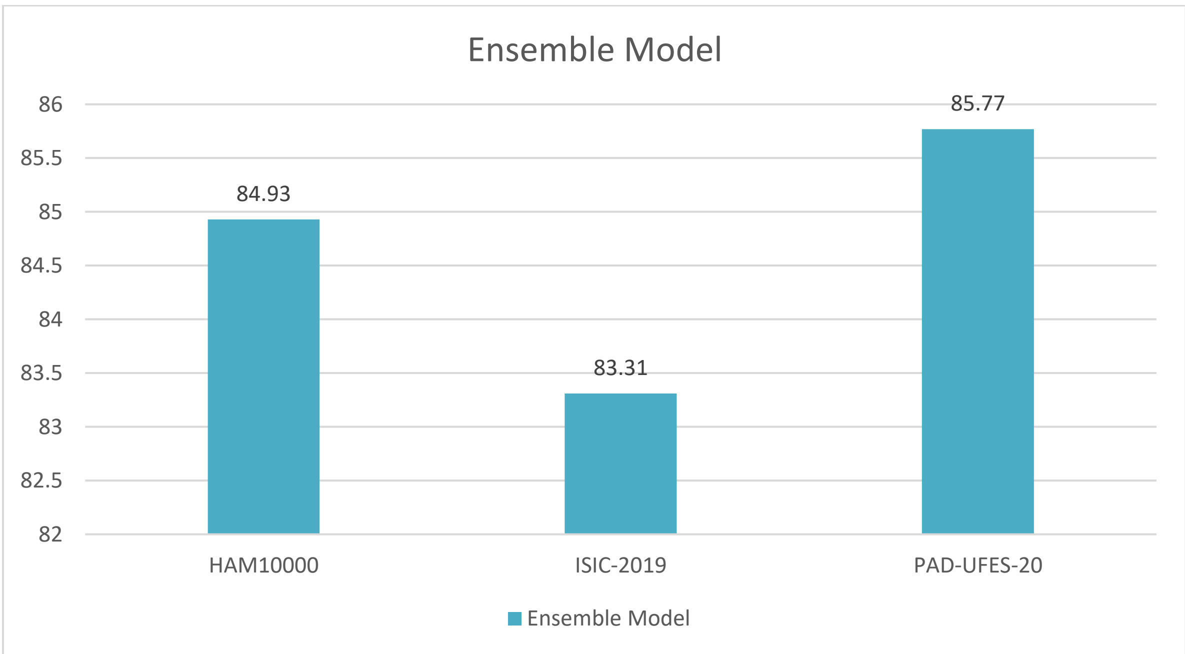


Figure 5.4.2.4: Ensemble Model Accuracy

5.4.3 Experimental Results of Proposed Model

In this experiment, the performance of the proposed teacher and student models with and without data augmentation was compared. The results indicated that data augmented models were able to achieve higher accuracy and show more stable loss curves during the training phase in every case. For instance, on the HAM10000 dataset, the teacher model with augmentation had an accuracy of 93.72%, and the model without augmentation had an 86.76% accuracy. The results of the student model for HAM10000 also showed an augmentation effect that was quite remarkable with a 93.41% accuracy rate against 84.60% accuracy without augmentation. On the ISIC-2019 dataset, the score of the teacher model with augmentation was 92.86% while the non-augmented model's score was 86.11%. For the student model on the ISIC-2019 dataset, the augmentation result was 92.59% and without augmentation it was 85.20% accurate. The best performance was achieved on the PAD-UFES-20 dataset with the augmented teacher model accuracy, which is 94.56% compared to 87.45% without the augmentation of the data. With augmentation, the student model on PAD-UFES-20 obtained 94.14%, while without augmentation, the same student model obtained 87.03%. These results point to the positive role of data augmentation in improving model performance, both in terms of accuracy and increasing model stability during the phases of learning, across all datasets.

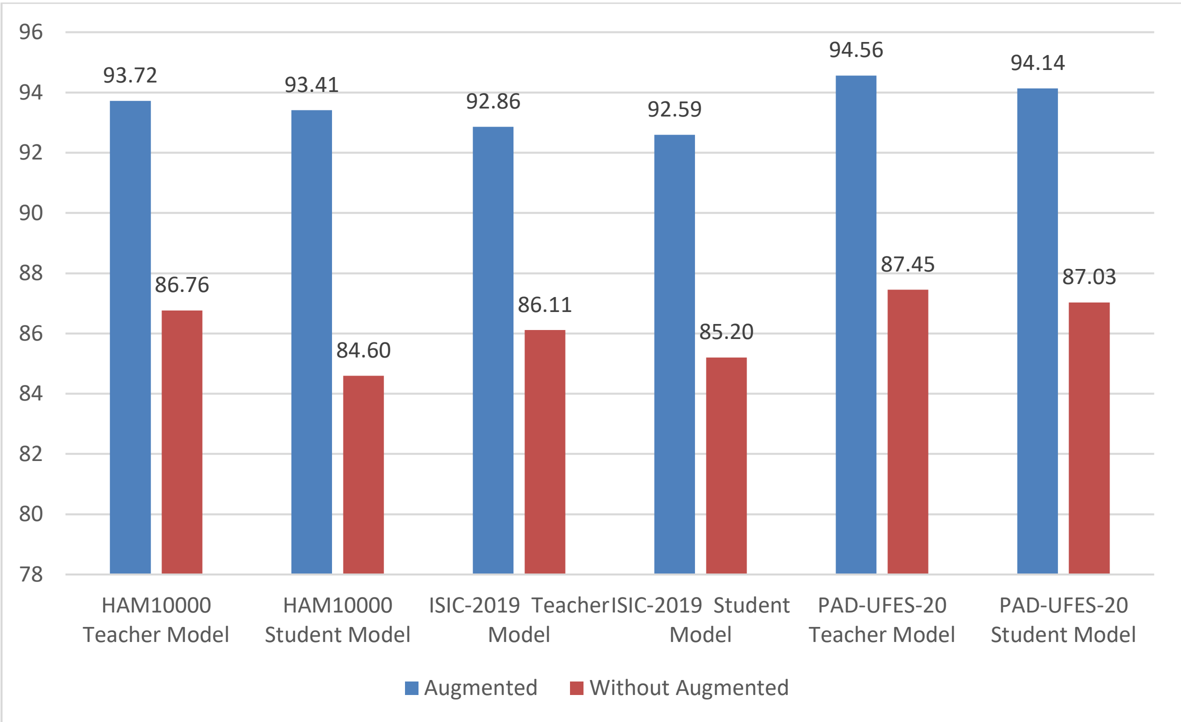


Figure 5.4.3.1: Proposed Model Accuracy (With and Without KD)

Classification Report:
In this section, we show the classification report of HAM10000, ISIC-2019 and PAD-UFES-20 datasets. To know the effect of data augmentation on model performance, we

compare the results of the model using the method of training without data augmentation and the result with data augmentation. The models were tested with essential metrics, namely accuracy, precision, recall and F1 score, to give an overview of how accurately they have classified skin lesions. The following tables summarize the performance of the proposed model for each dataset without and with augmentation for a detailed comparison of the results.

Table 5.4.3.1: Classification Report of the Teacher Model on HAM10000

Metric	**With Augmentation**							
	AK	**BCC**	**BKL**	**DF**	**MEL**	**NV**	**VASC**	**Overall**
Accuracy	63.64%	96.08%	92.73%	100.0%	70.27%	98.96%	100.0%	93.72%
Precision	95.45%	100.0%	86.44%	100.0%	84.78%	95.40%	100.0%	93.59%
Recall	63.64%	96.08%	91.89%	100.0%	70.27%	98.96%	100.0%	93.72%
F1-Score	76.36%	98.00%	89.08%	100.0%	76.85%	97.15%	100.0%	93.44%
Metric	**Without Augmentation**							
	AK	**BCC**	**BKL**	**DF**	**MEL**	**NV**	**VASC**	**Overall**
Accuracy	60.61%	84.31%	88.18%	75.00%	70.27%	93.00%	100.0%	86.76%
Precision	76.92%	95.56%	79.51%	100.0%	62.90%	91.76%	100.0%	86.57%
Recall	58.82%	84.31%	87.39%	75.00%	58.65%	93.83%	100.0%	86.76%
F1-Score	66.67%	89.58%	83.26%	85.71%	60.70%	92.79%	100.0%	86.55%

Table 5.4.3.2: Classification Report of the Student Model on HAM10000

Metric	**With Augmentation**							
	AK	**BCC**	**BKL**	**DF**	**MEL**	**NV**	**VASC**	**Overall**
Accuracy	63.64%	96.08%	90.91%	100.0%	70.27%	98.66%	100.0%	93.41%
Precision	95.45%	100.0%	86.21%	100.0%	82.98%	95.25%	100.0%	93.27%
Recall	63.64%	96.08%	90.91%	100.0%	70.27%	98.66%	100.0%	93.41%
F1-Score	76.36%	98.00%	88.50%	100.0%	76.10%	96.93%	100.0%	93.15%

Metric	Without Augmentation							
	AK	BCC	BKL	DF	MEL	NV	VASC	Overall
Accuracy	60.61%	84.31%	86.36%	75.00%	70.27%	90.76%	100.0%	84.60%
Precision	76.92%	82.69%	70.90%	100.0%	61.90%	91.58%	100.0%	84.78%
Recall	58.82%	84.31%	85.59%	75.00%	58.65%	90.76%	100.0%	84.60%
F1-Score	66.67%	83.50%	77.55%	85.71%	60.23%	91.17%	100.0%	84.55%

Table 5.4.3.3: Classification Report of the Teacher Model on ISIC-2019

Metric	With Augmentation								
	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	Overall
Accuracy	84.73%	98.99%	93.09%	100.0%	42.21%	100.0%	100.0%	100.0%	92.86%
Precision	98.71%	98.23%	89.86%	100.0%	75.00%	30.00%	100.0%	76.83%	94.50%
Recall	81.49%	98.99%	93.09%	100.0%	68.10%	100.0%	100.0%	100.0%	92.86%
F1-Score	89.28%	98.61%	91.45%	100.0%	71.38%	46.15%	100.0%	86.90%	93.29%
Metric	Without Augmentation								
	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	Overall
Accuracy	84.73%	88.98%	84.98%	100.0%	42.21%	100.0%	100.0%	100.0%	86.11%
Precision	86.07%	94.87%	83.48%	100.0%	52.11%	42.11%	100.0%	70.00%	87.85%
Recall	79.96%	88.91%	84.98%	100.0%	67.68%	100.0%	100.0%	100.0%	86.11%
F1-Score	82.90%	91.79%	84.23%	100.0%	58.89%	59.26%	100.0%	82.35%	86.66%

Table 5.4.3.4: Classification Report of the Student Model on ISIC-2019

Metric	With Augmentation								
	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	Overall
Accuracy	63.64%	96.08%	90.91%	100.0%	70.27%	98.66%	100.0%	100%	92.59%
Precision	95.45%	100.0%	86.21%	100.0%	82.98%	95.25%	100.0%	95.45%	94.19%

Recall	63.64 %	96.08 %	90.91 %	100.0 %	70.27 %	98.66 %	100.0 %	100%	92.59 %
F1-Score	76.36 %	98.00 %	88.50 %	100.0 %	76.10 %	96.93 %	100.0 %	97.67 %	93.08 %
Metric	**Without Augmentation**								
	MEL	**NV**	**BCC**	**AK**	**BKL**	**DF**	**VASC**	**SCC**	**Overall**
Accuracy	84.73 %	87.97 %	82.88 %	100.0 %	42.21 %	100.0 %	100.0 %	100.0 %	85.20 %
Precision	86.07 %	94.50 %	83.13 %	100.0 %	48.47 %	40.00 %	100.0 %	70.00 %	87.34 %
Recall	79.30 %	87.97 %	82.88 %	100.0 %	67.68 %	100.0 %	100.0 %	100.0 %	85.20 %
F1-Score	82.54 %	91.11 %	83.01 %	100.0 %	56.49 %	57.14 %	100.0 %	82.35 %	85.88 %

Table 5.4.3.5: Classification Report of the Teacher Model on PAD-UFES-20

Metric	**With Augmentation**						
	BCC	**ACK**	**NEV**	**SEK**	**SCC**	**MEL**	**Overall**
Accuracy	94.32%	96.05%	92.00%	96.00%	90.00%	100.0%	94.56%
Precision	96.51%	98.65%	92.00%	82.76%	100.0%	71.43%	95.05%
Recall	94.32%	96.05%	92.00%	96.00%	90.00%	100.0%	94.56%
F1-Score	95.40%	97.33%	92.00%	88.89%	94.74%	83.33%	94.67%
Metric	**Without Augmentation**						
	BCC	**ACK**	**NEV**	**SEK**	**SCC**	**MEL**	**Overall**
Accuracy	88.64%	88.16%	92.00%	84.00%	85.00%	60.00%	87.45%
Precision	91.76%	95.71%	76.67%	70.00%	100.0%	42.86%	88.83%
Recall	88.64%	88.16%	92.00%	84.00%	85.00%	60.00%	87.457%
F1-Score	90.17%	91.78%	83.64%	76.36%	91.89%	50.00%	87.86%

Table 5.4.3.6: Classification Report of the Student Model on PAD-UFES-20

Metric	**With Augmentation**						
	BCC	**ACK**	**NEV**	**SEK**	**SCC**	**MEL**	**Overall**
Accuracy	94.32%	96.05%	96.00%	96.00%	90.00%	60.00%	94.14%

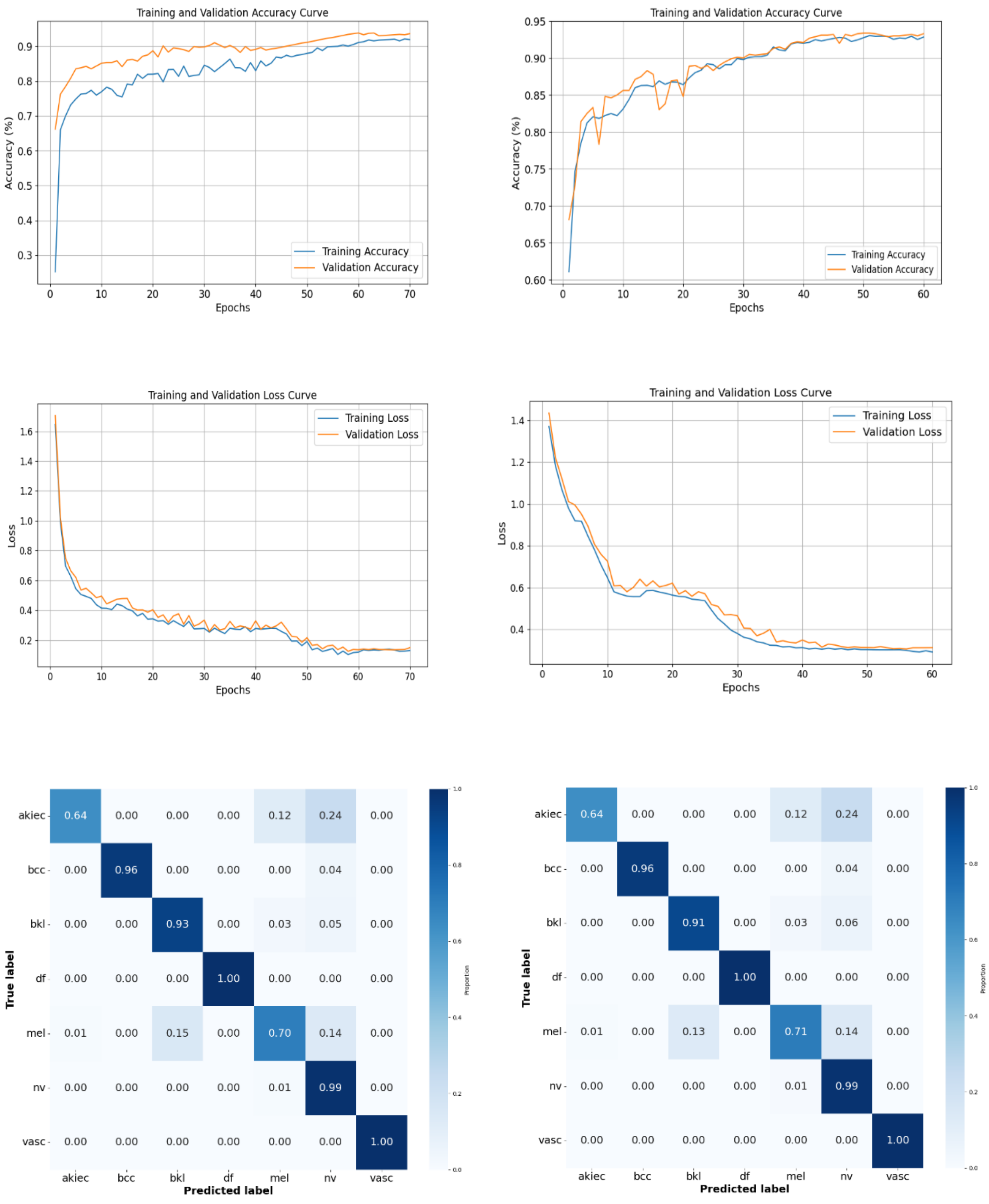
Precision	95.40%	98.65%	88.89%	82.76%	100.0%	75.00%	94.39%
Recall	94.32%	96.05%	96.00%	96.00%	90.00%	60.00%	94.14%
F1-Score	94.86%	97.33%	92.31%	88.89%	94.74%	66.67%	94.15%
Metric	**Without Augmentation**						
	BCC	**ACK**	**NEV**	**SEK**	**SCC**	**MEL**	**Overall**
Accuracy	88.64%	86.84%	92.00%	84.00%	85.00%	60.00%	87.03%
Precision	91.76%	97.06%	76.67%	65.62%	100.0%	42.86%	88.80%
Recall	88.64%	86.84%	92.00%	84.00%	85.00%	60.00%	87.03%
F1-Score	90.17%	91.67%	83.64%	73.68%	91.89%	50.00%	87.54%

From Table 5.4.3.1 to 5.4.3.6 we see that data augmentation helped in improve model performance for each of the three datasets of medical images. The augmented models compared to their non-augmented counterpart performed better, showing an increase between 6.75 to 8.81% in accuracy. HAM10000 showed the highest improvements with scores between 6.96% to 8.81%. PAD-UFES was followed by an improvement of 7.11%; ISIC-2019, an improvement of 6.75% - 7.39%. Student models showed higher response to augmentation than the teacher models, with an average increase of 8.04% vs. 6.94% for students and teachers, respectively. The augmentation technique improved all performance measures, including precision, recall and F1-scores for most individual disease classes. More importantly, augmentation helped to balance the dataset by improving performance on rare disease classes and retaining good results on common classes. These consistent improvements were observed across various datasets and model types, which is an indication of the need for data augmentation in the construction of reliable medical image classification systems.

Performance Visualization Analysis:

In this part, a visual comparison was made between the augmented and non-augmented versions of the proposed model. The results show higher accuracy and more stable loss curves while training using the augmentation technique. In contrast, the version without augmentation had uncontrollable training behavior, still had lower accuracy, emphasizing the power of the data augmentation technique. These findings provide a lot of evidence that the applied augmentation helped improve the model performance, improving stability and accuracy during the training phase on the datasets.

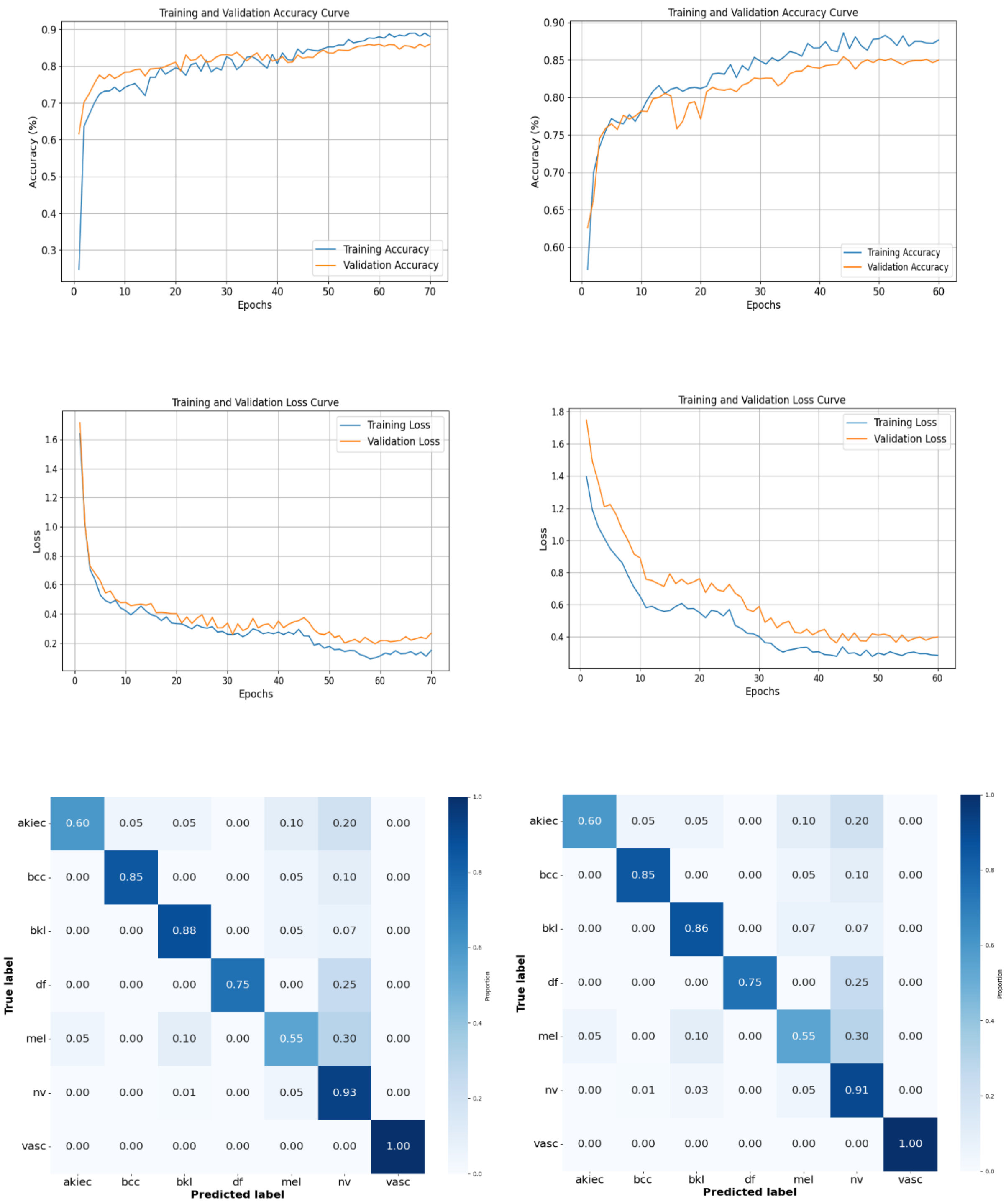
HAM10000 Dataset (Augmented):



(A) (B)

Figure 5.4.3.2: Accuracy, Loss, and Confusion Matrix of (A)Teacher and (B)Student Models on Augmented HAM10000 Dataset

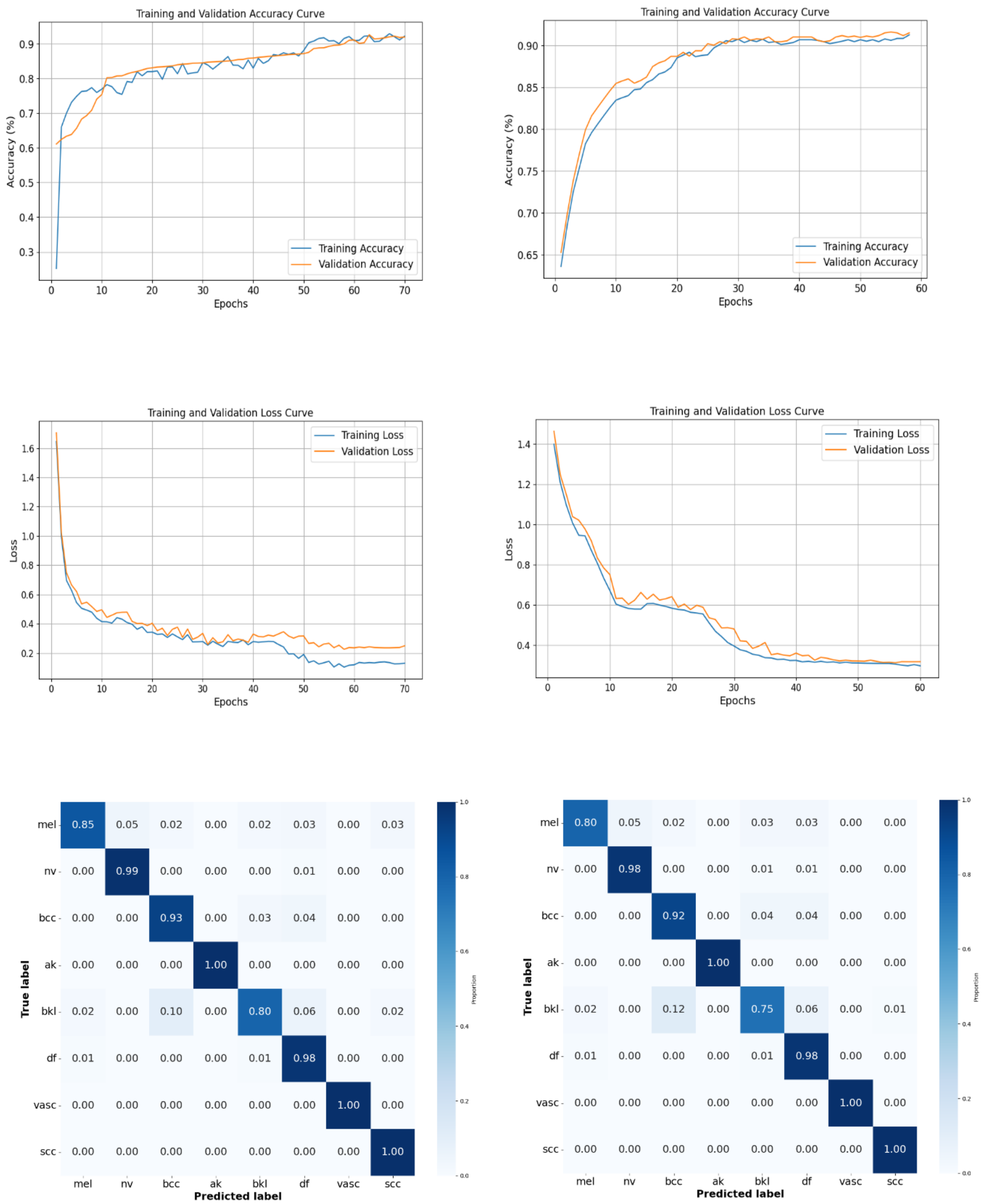
HAM10000 Dataset (Without-Augmentation):



(A) (B)

Figure 5.4.3.3: Accuracy, Loss, and Confusion Matrix of (A)Teacher and (B)Student Models on non-Augmented HAM10000 Dataset

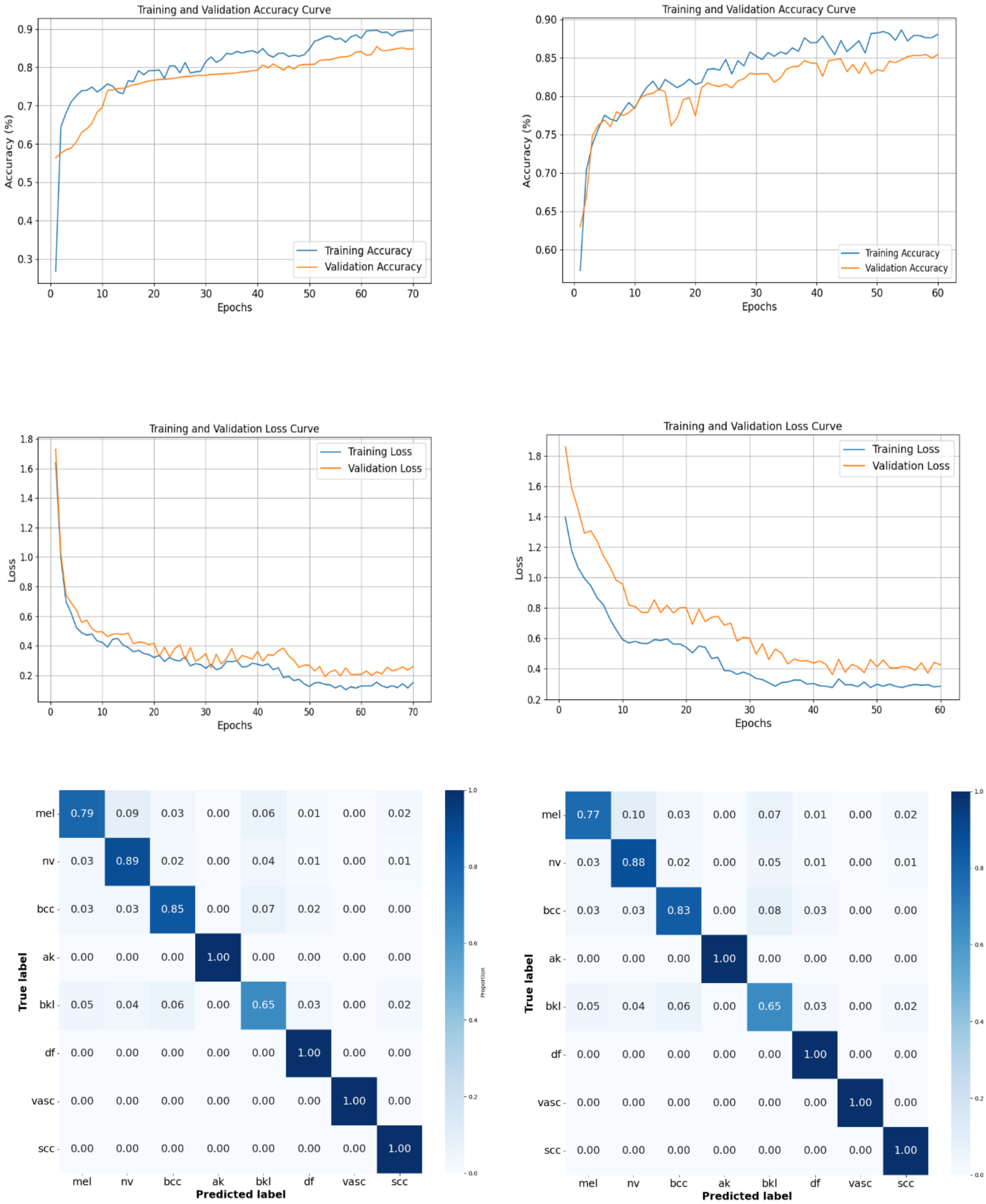
ISIC-2019 Dataset (Augmented):



(A) (B)

Figure 5.4.3.4: Accuracy, Loss, and Confusion Matrix of (A)Teacher and (B)Student Models on Augmented ISIC-2019 Dataset

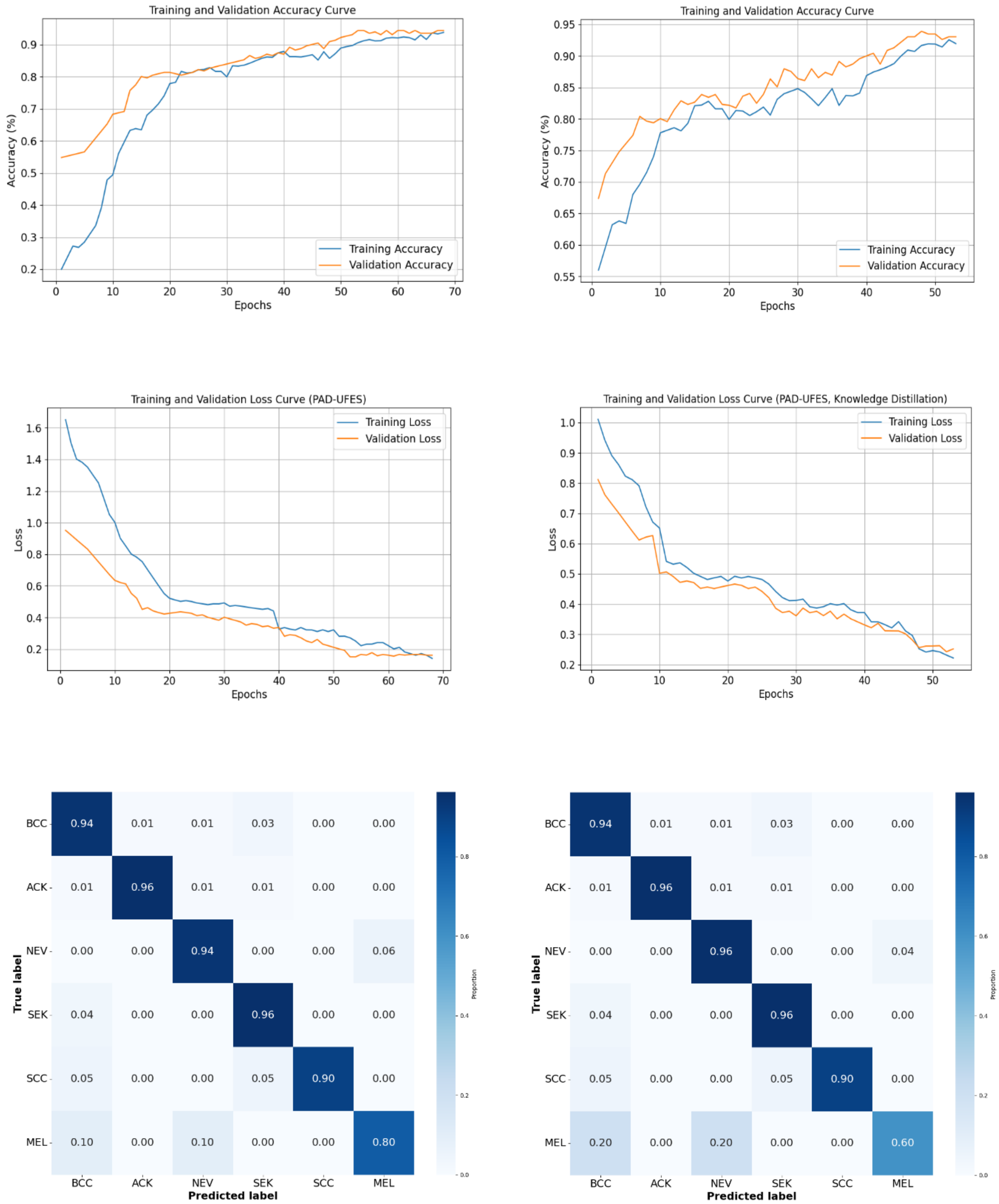
ISIC-2019 Dataset (Without-Augmentation):



(A) (B)

Figure 5.4.3.5: Accuracy, Loss, and Confusion Matrix of (A)Teacher and (B)Student Models on non-Augmented ISIC-2019 Dataset

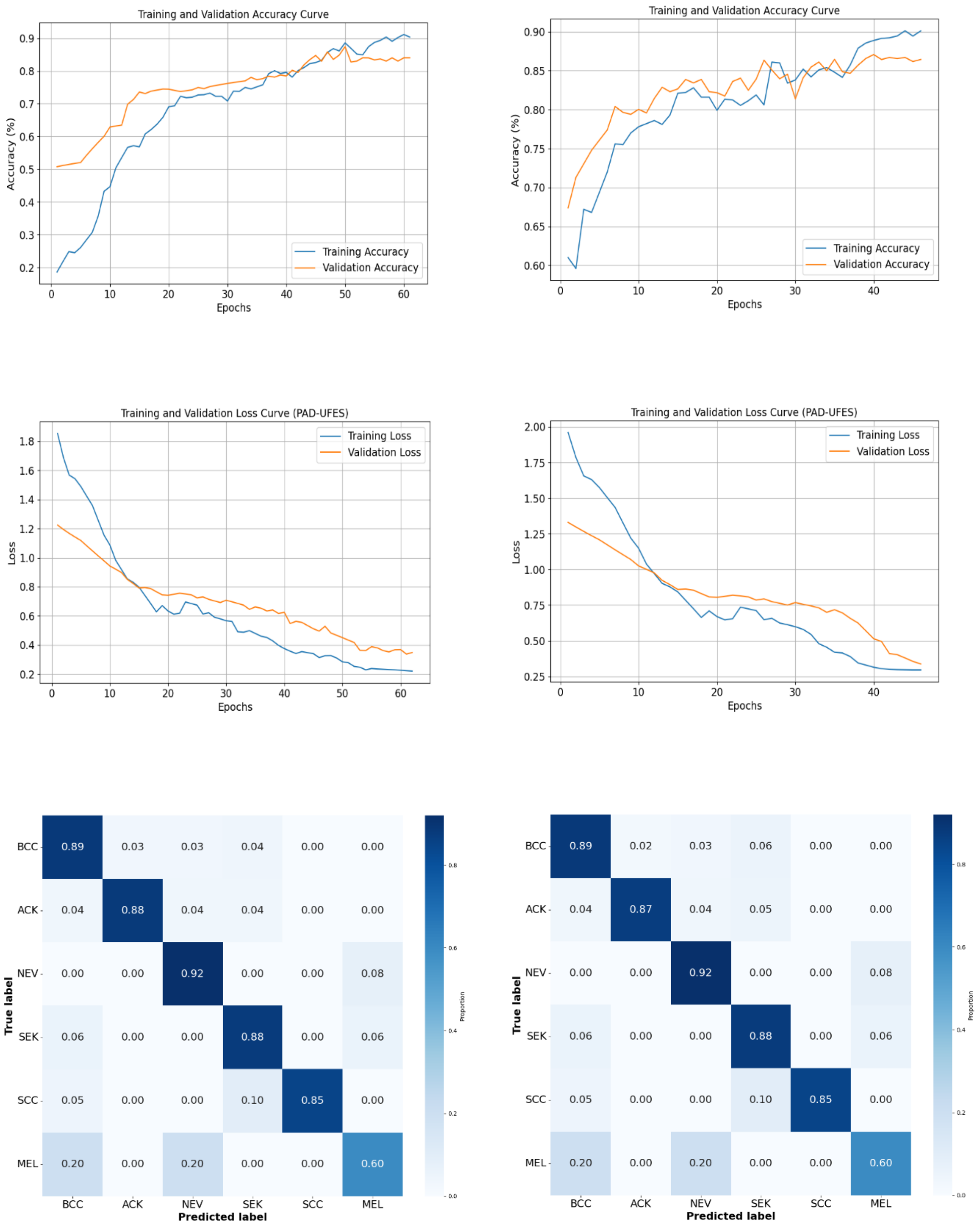
PAD-UFES-20 Dataset (Augmented):



(A) (B)

Figure 5.4.3.6: Accuracy, Loss, and Confusion Matrix of (A)Teacher and (B)Student Models on Augmented PAD-UFES-20 Dataset

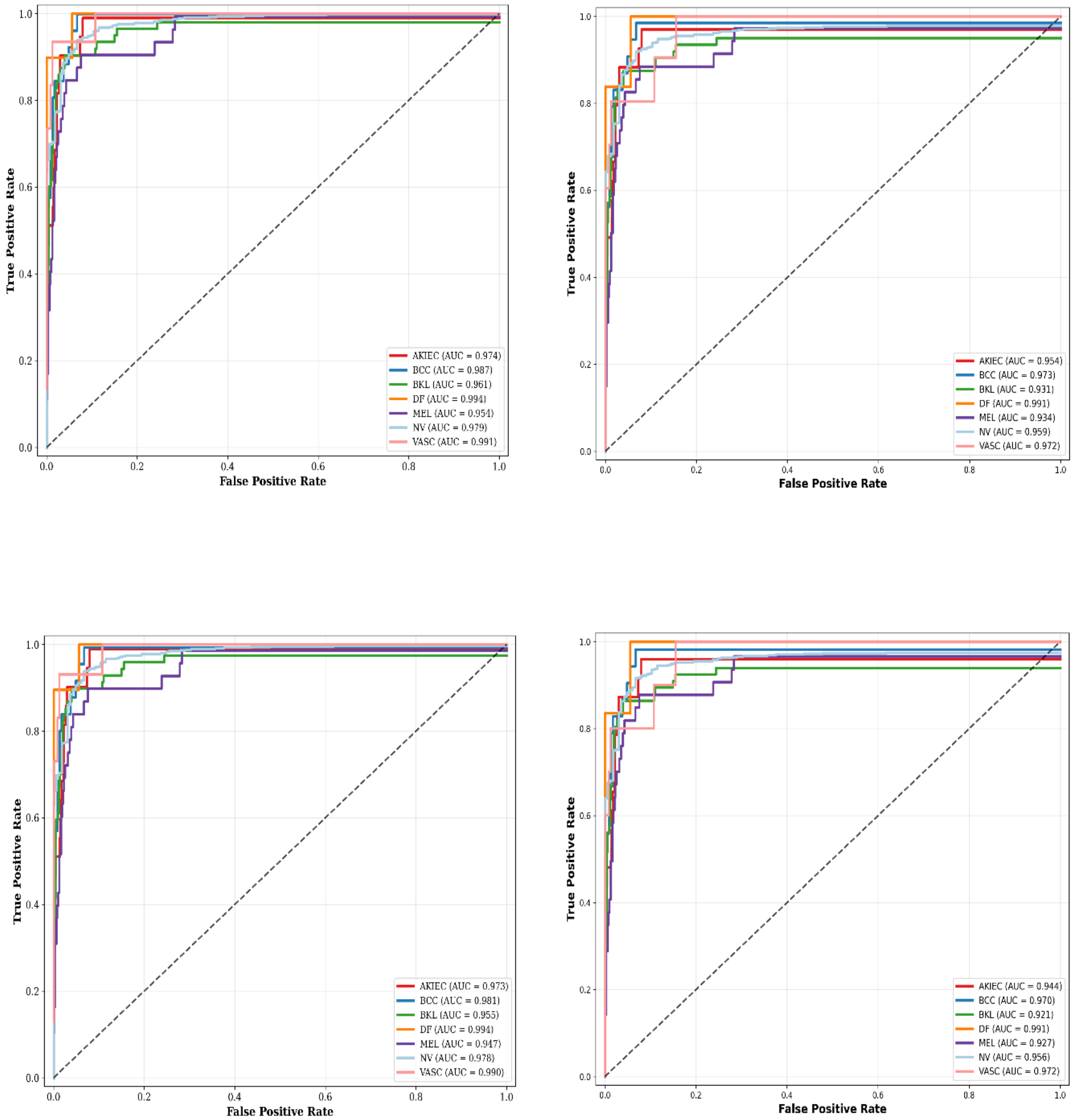
PAD-UFES-20 Dataset (Without-Augmentation):



(A) (B)

Figure 5.4.3.7: Accuracy, Loss, and Confusion Matrix of (A)Teacher and (B)Student Models on non-Augmented PAD-UFES-20 Dataset

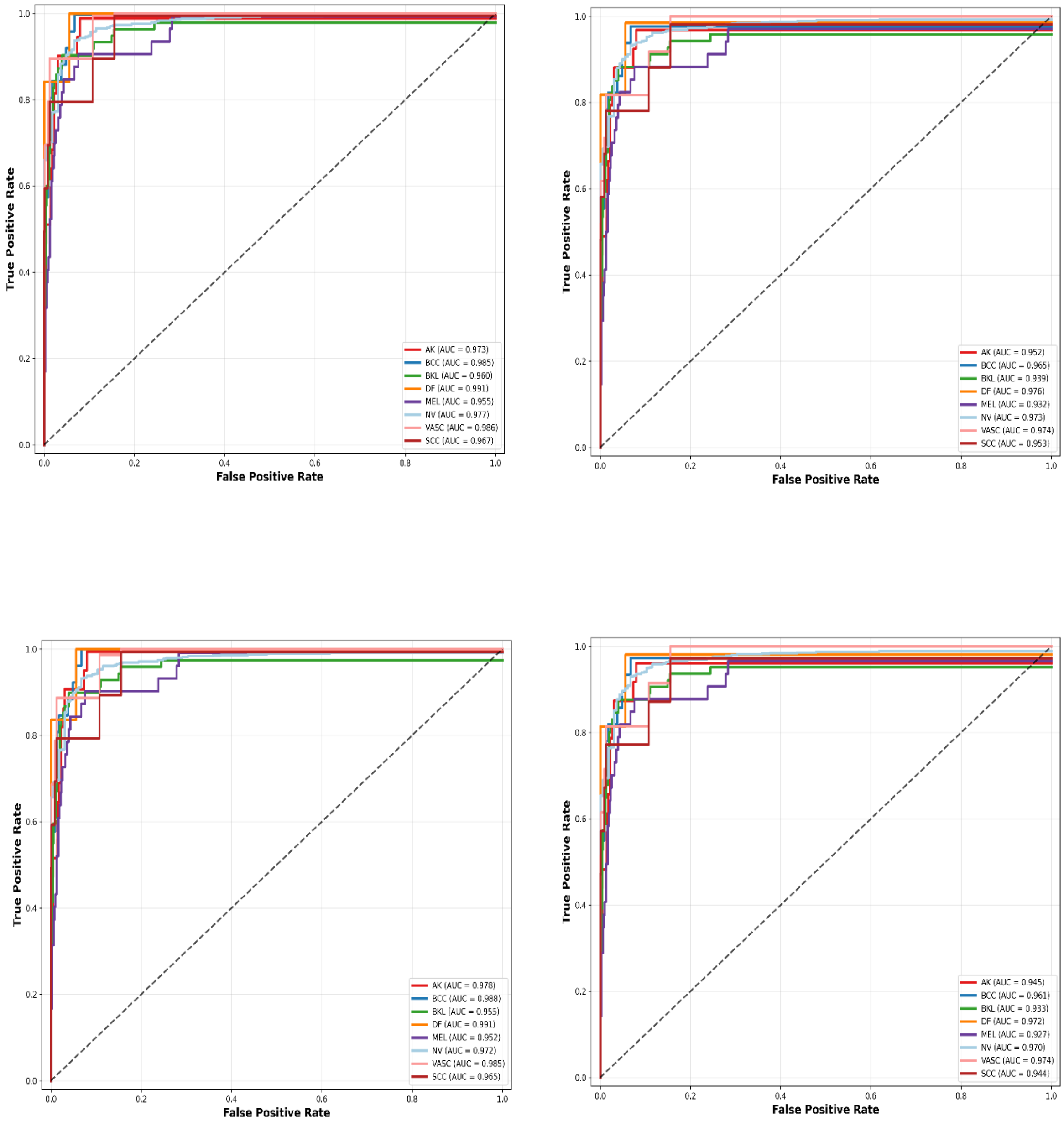
HAM10000 Dataset (ROC Curve):



(A) (B)

Figure 5.4.3.8: ROC Curve of Teacher and Student Models on (A)Augmented and (B)Non-Augment of HAM10000 Dataset

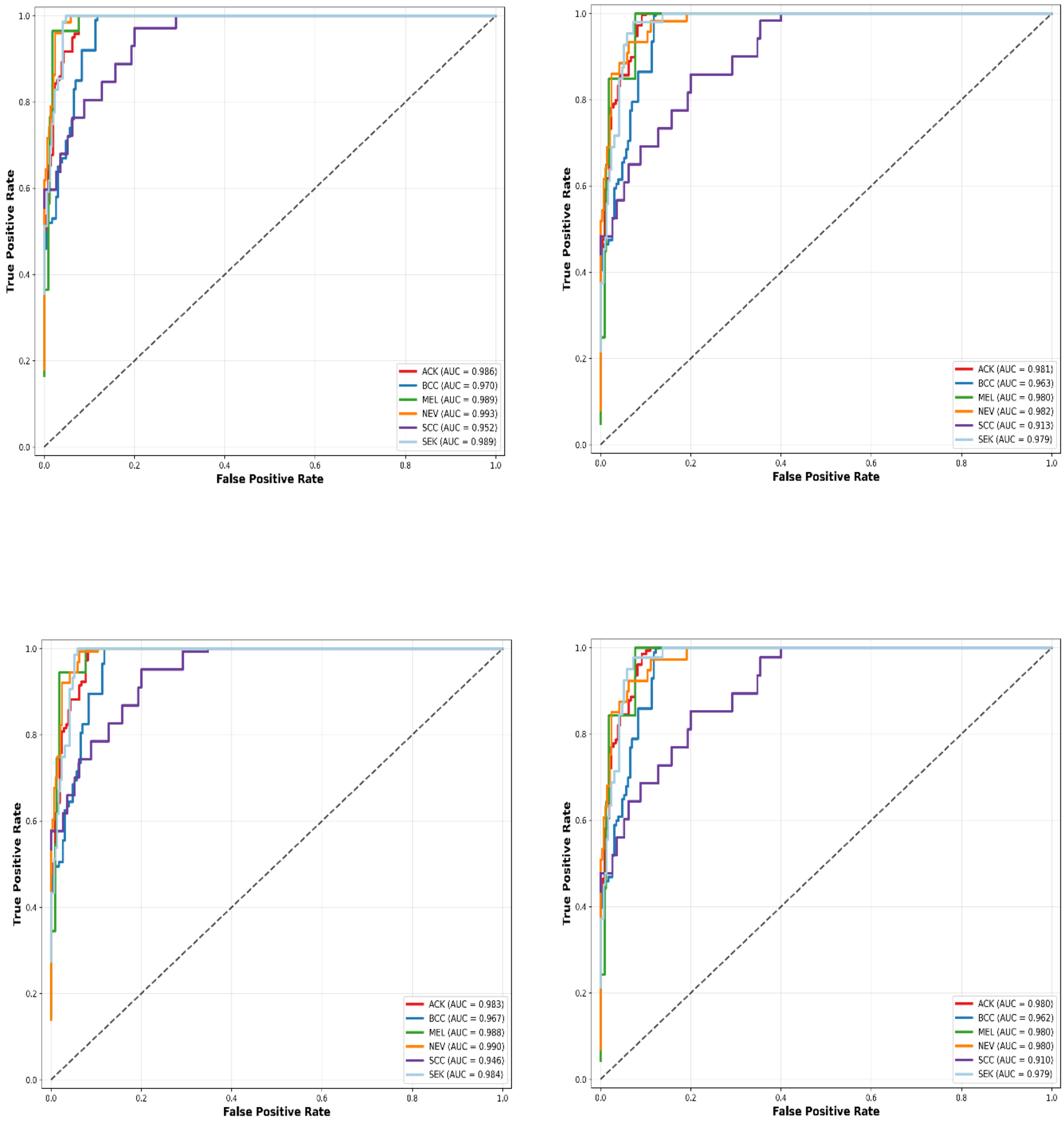
ISIC-2019 Dataset (ROC Curve):



(A) (B)

Figure 5.4.3.9: ROC Curve of Teacher and Student Models on (A)Augmented and (B)Non-Augment of ISIC-2019 Dataset

PAD-UFES-20 Dataset (ROC Curve):



(A) (B)

Figure 5.4.3.10: ROC Curve of Teacher and Student Models on (A)Augmented and (B)Non-Augment of PAD-UFES-20 Dataset

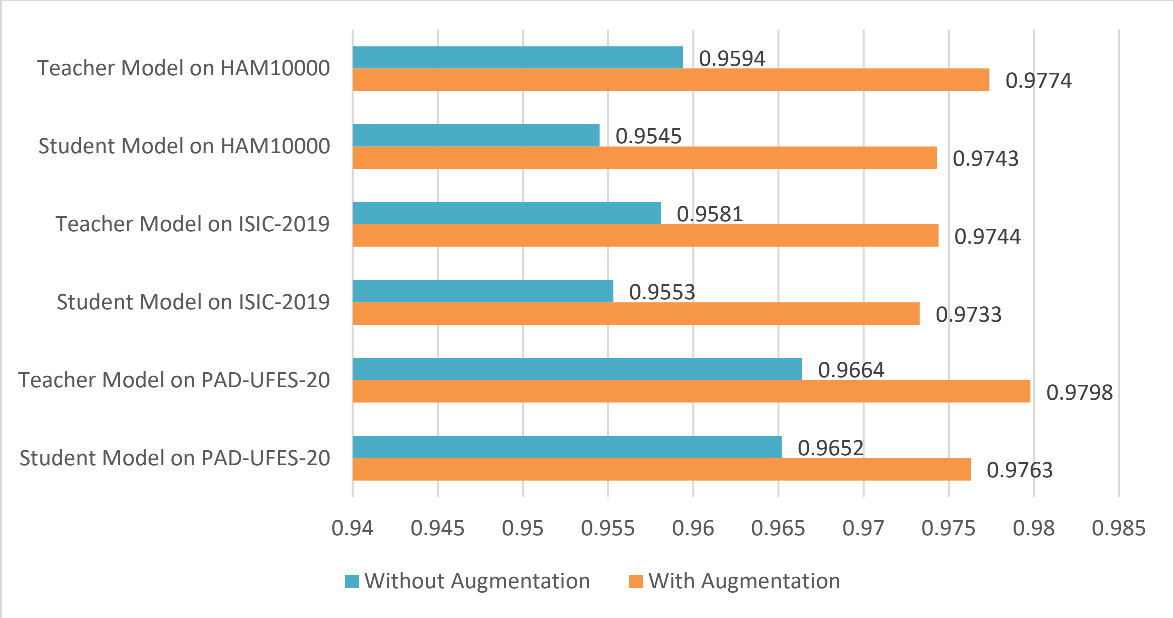


Figure 5.4.3.11: AUC Comparison of Teacher and Student Models on HAM10000, ISIC-2019, and PAD-UFES-20 Datasets with and without Augmentation

5.4.4 Ablation Study

This section examines the influence of some key components of the proposed model: LoRA, multi-axis attention, and knowledge distillation over three datasets: HAM10000, ISIC-2019, and PAD-UFES-20.

5.4.4.1 Impact of LoRA and Multi-Axis Attention

Impact of LoRA: LoRa reduces the model parameters and makes the training process much efficient without reducing any significant accuracy. It greatly reduces the training time while keeping the performance.

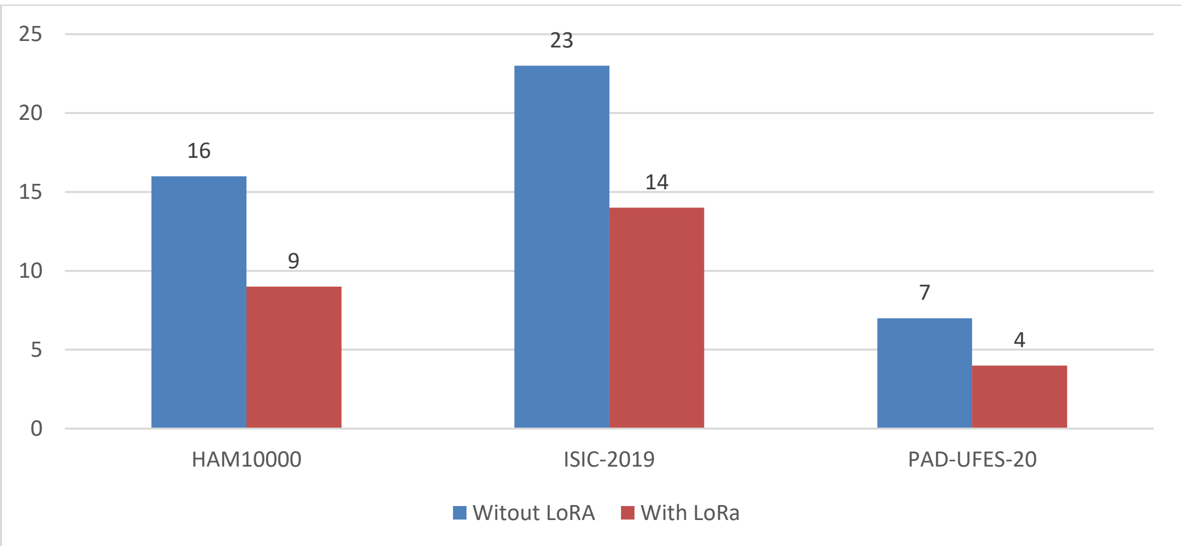


Figure 5.4.4.1.1: Impact of LoRA on Training Time on Hours

Impact of Multi-Axis Attention: This section examines the impact of multi-axis attention on the teacher model. By adding local and global attention, the model attention was more complex, capturing spatial and auditory relationships better enabling it to focus on relevance features, texture and boundary patterns on skin lesions.

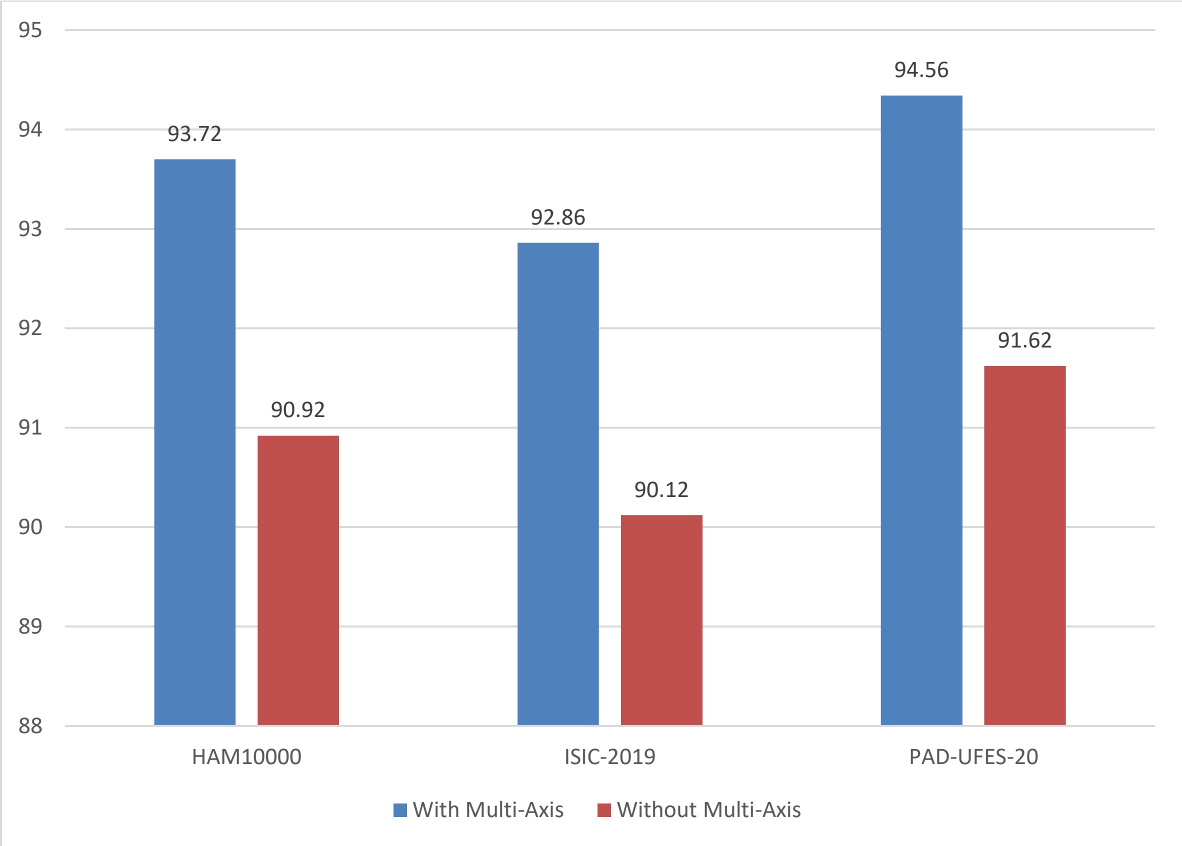


Figure 5.4.4.1.2: Impact of Multi-Axis Attention on Accuracy with Teacher Model

5.4.4.2 Impact of Knowledge Distillation

Knowledge distillation is a method to transfer knowledge from a larger teacher model, which is built to be more accurate to a smaller but more efficient student model. This technique minimizes the size of the model and computation requirements with a good level of accuracy.

In this research, the student model TinyViT, which has only 5 million parameters, is used. The following chart shows the performance of the student model with and without knowledge distillation from the teacher model. Despite its smaller size, the student model showed excellent knowledge transfer and a significant boost in efficiency with relatively little loss in accuracy, in comparison to the teacher model. This shows the power of knowledge distillation to maintain performance while saving resources at the same time.

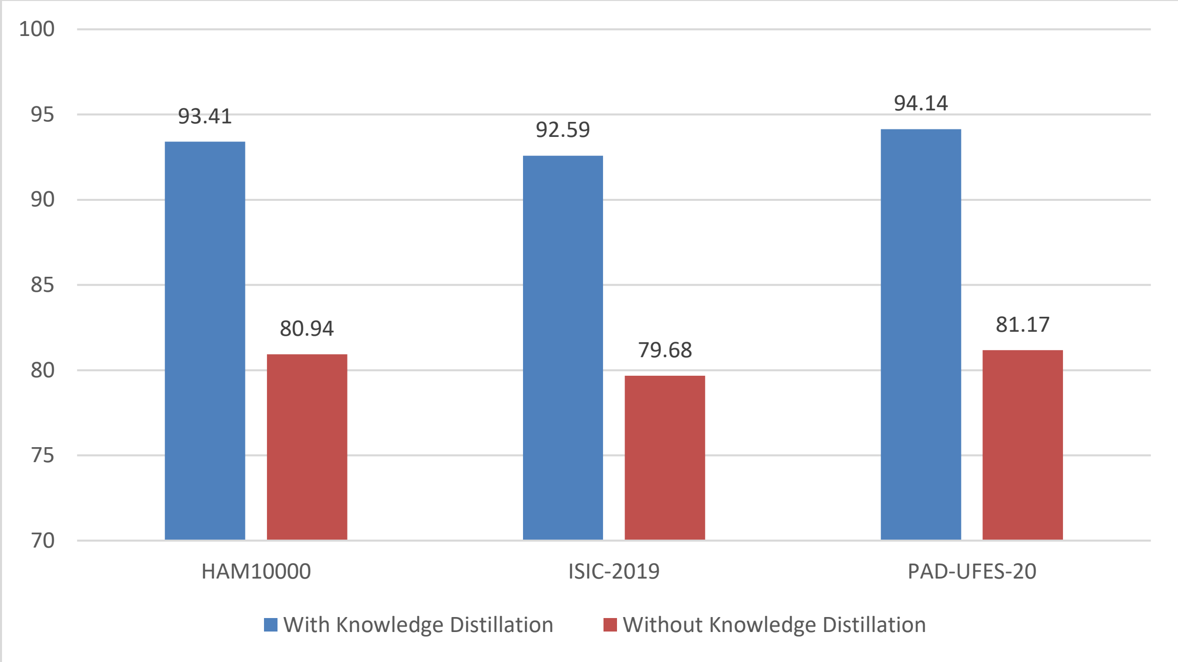


Figure 5.4.4.2: Impact of Knowledge Distillation on Student Model

5.4.5 Comparison with Existing Research

Table 5.4.5 compares the accuracy of our proposed model with other state-of-the-art candidate models. Among them, Saha et al. [29] and Liu et al. [30] adopted a knowledge distillation strategy in which they attained accuracy up to 93.24% and 88.5% by taking advantage of a pretrained teacher model, such as ResNet50 and MobileNetV2 and EfficientNetB0 for the student model, respectively. Comparatively, our proposed model that relied on self-supervised learning, multi-axis attention, and LoRA achieved an accuracy score of 94.56% and 94.14% on the teacher model and student model, respectively. This shows the superiority of our approach, from being better than both models in terms of classification performance and effectiveness.

Table 5.4.5: Comparison of Proposed Model with Existing Research

Author (s)	Methodology	Best Accuracy	Year
Kaur et al. [11]	End-to-end system with segmentation (ACNN), contrast enhancement (MCAN), hair removal, and data augmentation for N-DCNN	93.40% (N-DCNN)	2025
Gururaj et al. [12]	Transfer learning with lesion segmentation, hair removal, and class balancing using oversampling, undersampling for ResNet50 and DenseNet169	91.2% (DenseNet169)	2023

Tuncer et al. [15]	Lightweight CNN (TurkerNet) with depth-wise separable convolutions for binary classification	92.12% (TurkerNet)	2024
Vega-Huerta et al. [17]	Custom 19-layer CNN with normalization, pooling, dropout, and data augmentation for multi-class classification	94% (Custom CNN)	2025
Islam et al. [18]	Benchmarking of multiple architectures including ConvNeXt V2, SwinV2, ViT, with standard preprocessing	93.2% (ConvNeXt V2)	2024
Flosdorf et al. [19]	Evaluation of pre-trained ViT-L16 and ViT-L32 with initial preprocessing	92.79% (ViT-L16)	2024
Zhang et al. [20]	Modular ViT (DermViT) with Dermoscopic Context Pyramid, hierarchical attention, and feature gating	86.12% (DermViT)	2025
Leema et al. [22]	Lightweight Multi-Scale ViT (LMS-ViT) with multi-scale attention and compact feature fusion for real-time classification	90% (LMS-ViT)	2025
Pacal et al. [24]	MetaFormer-based model (CAFormer) with focal self-attention, data augmentation, and normalization	95.01% (CAFormer)	2025
Ozdemir et al. [25]	Hybrid framework integrating ConvNeXtV2 with focal self-attention modules and sliding window dual-path architecture	93.60% (Hybrid ConvNeXtV2)	2025
Dagnaw et al. [26]	Hybrid model combining ViTs, CNNs (ResNet50, etc.), and Swin Transformers with SMOTE balancing and interpretability tools	88.8% (ResNet50)	2024
Tembhurne et al. [27]	Ensemble of ML and DL models with feature extraction and classification	93% (Ensemble ML, DL)	2023
Saha et al. [29]	Two-phase knowledge distillation using pre-trained teachers (ResNet50, etc.) to supervise lightweight students with data augmentation	93.24% (Custom Student Model)	2025
Liu et al. [30]	Relation-aware mutual learning between identical student models (MobileNetV2, EfficientNetB0) with weighted loss and augmentation	88.5% (EfficientNetB0)	2024

Rahman et al. [35]	Unsupervised k-means with histogram-based features, ESRGAN super-resolution, PCA+t-SNE dimensionality reduction, and augmentation	90.50% (Unsupervised k-means)	2025
Proposed Model	**Self-supervised+Knowledge Distillation+LoRA+Multi-Axis Attention+**	**94.56% (Teacher) 94.14% (Student)**	**2025**

5.5 Results and Discussion

The proposed teacher model with self-supervised learning (SSL) and multi-axis attention achieved outstanding performance for all the data sets with 93.72%, 92.86%, and 94.56% accuracy for HAM10000, ISIC-2019, and PAD-UFES-20, respectively. The student model benefited from knowledge distillation from the teacher model, obtaining 93.41%, 92.59%, and 94.14% in the same datasets, while only using 5 million parameters. Both models outperformed the ensemble model and several well-known pre-trained models, such as ViT and SwinV2-b, demonstrating that the combination of SSL, multi-axis attention, LoRA and knowledge distillation is a powerful way for improving performance and efficiency.

Table 5.5: Result Analysis of All Experimental Models

Model	**Parameters**	**Accuracy**		
		HAM100000	**ISIC2019**	**PAD-UFES-20**
ViT	86.57m	87.43%	86.90%	87.87%
SwinV2-b	88m	87.13%	86.42%	87.03%
ConvNeXt-B	88.6m	86.93%	86.54%	87.45%
ResNet101	44.55m	86.63%	86.31%	86.19%
DenseNet201	18.57m	86.43%	86.39%	86.61%
EfficientNetB4	19m	85.03%	83.90%	85.77%
TinyVit	5m	80.94%	79.68%	81.17%
EfficientNetB1	7.84m	76.65%	75.93%	76.99%
MobileNetV3	5.4m	76.45%	75.37%	76.15%

ResNet18	11.69m	74.45%	73.24%	74.90%
DenseNet121	7.98m	77.15%	75.85%	77.82%
Ensemble Model	16m	84.93%	83.31%	85.77%
Proposed Teacher Model	96m	93.72%	92.86%	**94.56%**
Proposed Student Model Without (KD)	5m	80.94%	79.68%	81.17%
Proposed Student Model	5m	93.41%	92.59%	**94.14%**

5.6 Explainable AI to Interpret Model Predictions

In this part, we focus on the use of explainable AI techniques to gain an understanding of how the proposed model is making its predictions. We focus on Grad-CAM, which visualizes regions of an image that the model thinks are most critical to classify skin lesions. Grad-CAM produces heatmaps that are used to emphasize these critical regions, as well as create transparency into the decision-making process for the model. The following figure shows some Grad-CAM outputs from some selected images, showing how the model captures features it uses to make predictions. This interpretation helps to build trust and understanding of the behavior of the model.

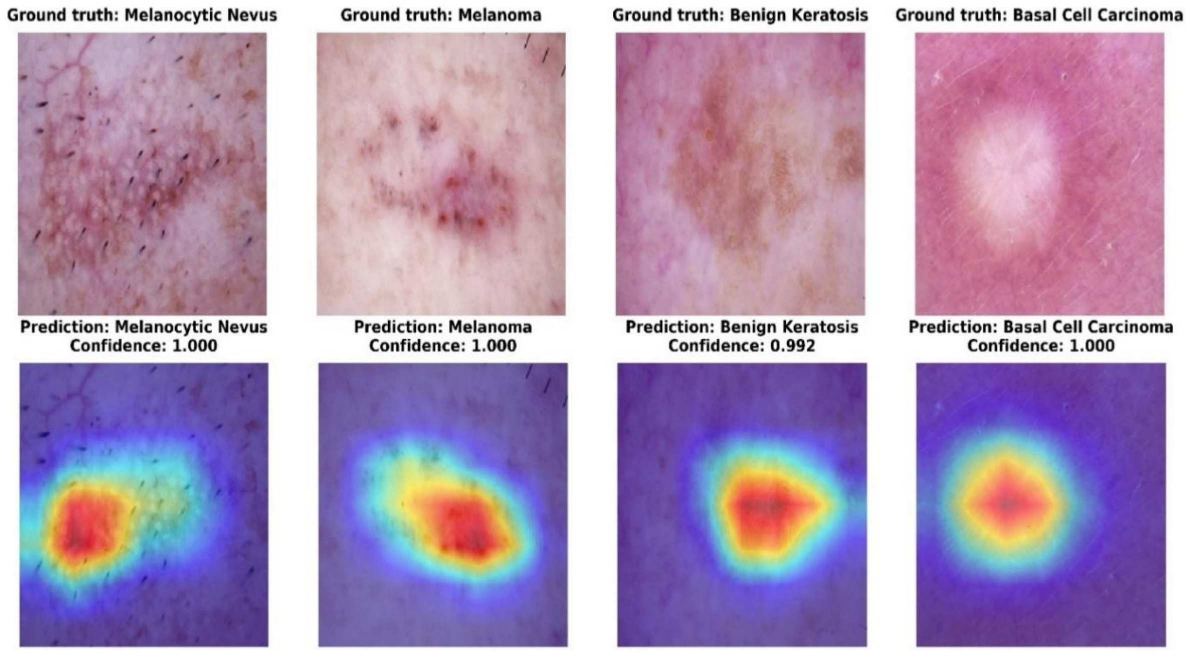


Figure 5.6.1: Grad-CAM Visualization of Model Prediction (1)

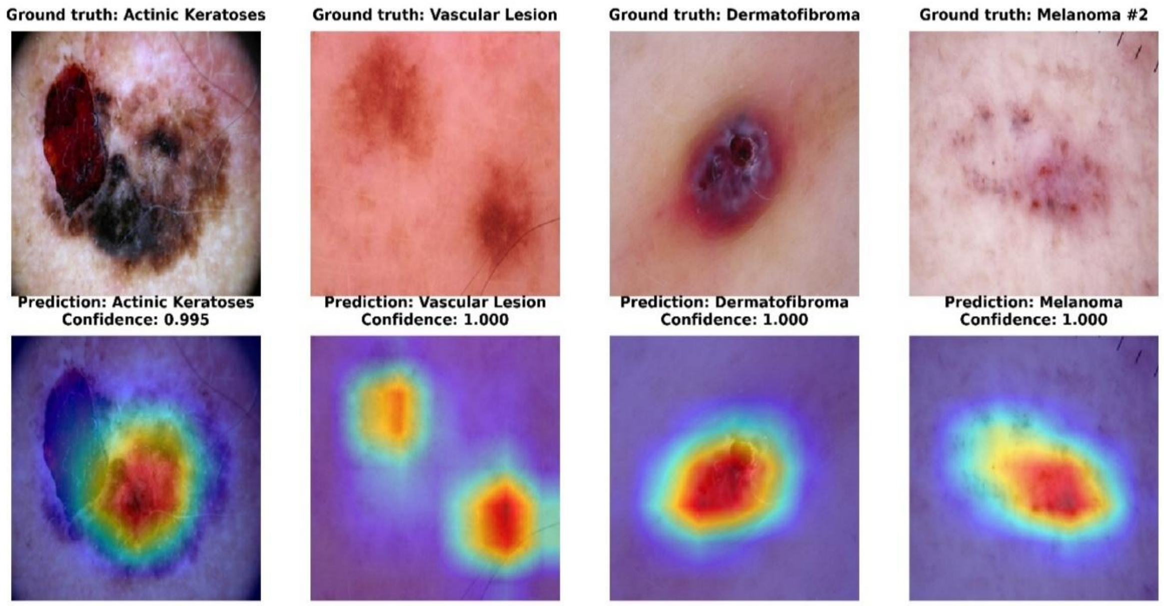


Figure 5.6.2: Grad-CAM Visualization of Model Prediction (2)

5.7 Summary

In this chapter, the performance of the proposed teacher and student models is evaluated in classifying skin lesions on three datasets: HAM10000, ISIC-2019 and PAD-UFES-20. Compared to several pre-trained models and the ensemble model, the teacher model had the highest accuracy, with 93.72%, 92.86%, and 94.56% accuracy on HAM10000, ISIC-2019, and PAD-UFES-20. The student model, based on the knowledge distillation approach, gets an accuracy of 93.41%, 92.59%, and 94.14% with just 5 million parameters, indicating that the model is effective. Techniques such as self-supervised learning, multi-axis attention, LoRA and knowledge distilling improved the performance of both models, which led to a trade-off between performance and computational efficiency, suitable for clinical application.

Chapter 6

Engineering Standards and Design Challenges

This chapter examines the engineering standards and design challenges of the skin cancer detection system that deal with compliance with software, hardware, and communication standards, as well as the impact of the system on society, the environment and sustainability. It is also eligible to cover project management, financial analysis, complex problem solving in engineering, and a summary to evaluate the feasibility and wider implications of the proposed solution.

6.1 Compliance with the Standards

In this section, we describe the engineering standards and regulations that were followed for the research, and the methodologies that have been used for the construction of the model for the classification of skin lesions. Ensuring that relevant standards are adhered to enables continuation of the quality, efficiency and dependability of the project. The standards discussed here is software, hardware and communication guidelines relevant for the development of machine learning models for medical image analysis.

6.1.1 Software Standards

The software used in this research follows well-established guidelines for deep learning model development. Python 3.8 was used as the programming language with PyTorch & TorchVision as the main frameworks for Deep learning. PyTorch's flexibility allowed for the validation of complicated architectures, and TorchVision allowed for efficient tools for image processing and augmentation. The software was also based upon modularity standards for ease of maintenance, scalability.

6.1.2 Hardware Standards

The model's training system was composed of an Nvidia GeForce RTX 3060 graphics card and an Intel Core i7 processor with 16 gigabytes of RAM. This hardware configuration permitted the training on massive medical datasets in an efficient way, taking into account the requirements of the deep learning algorithm. The hardware is sufficient for research purposes and can possibly be expanded for the limited-resource environment for possible deployment.

6.1.3 Communication Standards

Although there were no specific communication requirements on this research, the model's architecture is outlined with future web-based or application-based implementations in mind. This flexibility helps define that the model is adaptable to be integrated into clinical systems or clinical applications according to their need.

6.2 Impact on Society, Environment and Sustainability

6.2.1 Impact on Life

The proposed deep learning model of the classification of skin lesions, in particular it can help in the early detection of skin cancer, and most notably of skin cancer. With advancements in speed and accuracy of diagnosis, it means healthcare providers are able to intervene sooner in the illness and have a healthier outcome and increase survival. The system also offers a potential for wider access to quality diagnostics, especially in less served areas, where advanced biological screening for skin cancer can be made accessible to a greater number of individuals, and leads to overall betterment in the public health structure.

6.2.2 Impact on Society & Environment

By making diagnosis more automated, the system frees up the time of healthcare professionals so they can concentrate on more complex cases. Not only does this improve the efficiency of the medical workflows, but it also decreases the time between diagnosis and treatment to the benefit of the patient. Environmentally, the model's digital nature means that when you use it, you need less paper and lower physical resources. Moreover, the model also optimizes the use of healthcare resources as it improves diagnostic accuracy, thereby resulting in a reduction of unnecessary tests and treatments, leading to more sustainable forms of medical practice.

6.2.3 Ethical Aspects

The use of AI in this research is subject to significant moral considerations, to ensure guidance that is not loaded or untrustworthy.

- **Data Privacy and Security:** The privacy of patient care is of paramount importance here and we adhere to several accountability standards, including GDPR and HIPAA standards, by anonymizing and encrypting medical images and securely storing them, ensuring trust from healthcare providers
- **Bias and Fairness:** To avoid bias, the model is trained on a diverse dataset and regularly audited to ensure consistent accuracy across different demographics for fair and equitable results.

- **Transparency and Explainability:** We solve the "black box" problem by providing interpretable outputs, helping to build trust that leads to meaningful clinical decisions.
- **Accountability:** Decent rules needed to establish a sense of responsibility should mistakes be made, whether it is the developer, the healthcare provider, or even the system itself, accountability is provided and the corrective step may be undertaken.
- **Regulatory Compliance:** The system is in compliance with the healthcare standards and regulatory compliance is appropriately tracked to ensure that the system is up to date with the current standards.
- **Employment Effects:** AI will be used to supplement physicians and to reposition their more complex assignments as well as keep them in their important role.
- **Informed Consent and User Control:** The use of data is disclosed to the patient, and consumers have user control over data use and retailing, with no intervening exit.

6.2.4 Sustainability Plan

In order for the model to be sustainable long term, it will be important to regularly update it with new trends and new medical information, and keep the model accurate and relevant by consistently retraining and by integrating new data. Financially, the model should be developed in a scalable, accessible way for healthcare providers in resource-limited environments. Furthermore, the environmental footprint of the system can be minimized through optimizing energy consumption in training and deployment, energy-efficient infrastructure, and consideration of renewable energy options to power the model's operation.

6.3 Project Management and Financial Analysis

Effective project management ensured that the research was carried out on time, and there were clear objectives for each phase of the research. The project underwent a progressive plan, from the preparation of the data set, the training of the model, the evaluation, and the analysis. Table 6.3.1 and Table 6.3.2 represent the initial budget and alternate budget cost analysis, respectively.

Table 6.3.1: Initial Budget Table

Serial	Item	Description	Estimated Cost (BDT)
1	Hardware	Intel Core i7 13700 Processor	36000
		Nvidia RTX 3060 GPU	46000

		Ram 16GB	12000
		SSD 512GB	7600
2	Software	PyCharm	30000
3	Service	Development libraries, GitHub Pro	15000
4	Research Resource	Academic papers and articles	20000
5	Operating Costs	Internet, Electricity	10000
6	Miscellaneous Equipment	Basic Peripherals	5000
		Total	171600

Table 6.3.2: Alternate Budget Table

Serial	**Item**	**Description**	**Estimated Cost (BDT)**
1	Hardware	Cloud computing with Colab Pro	24000
2	Software	Open source IDE	0
3	Service	Open source libraries, GitHub	5000
4	Research Resource	Free Academic papers and articles	0
5	Operating Costs	Internet, Electricity	10000
6	Miscellaneous Equipment	Basic Peripherals	500
		Total	39500

6.4 Complex Engineering Problem

The research in this thesis, "An Approach to Improve the Efficiency of Deep Learning Framework for Skin Cancer Classification Utilizing Dermoscopy Images," is addressing a Complex Engineering Problem. It covers advanced machine learning and biomedical engineering with a high level of analytical skills and the ability to manage imbalanced datasets and build creative computational methods for efficient classification of medical images.

6.4.1 Complex Problem Solving

This section, provide a mapping with problem-solving categories in Table 6.4.1.1.

Table 6.4.1.1: Mapping Complex Engineering Problem

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	✓	×	✓	✓

Mapping with Knowledge Profile:

This section aims at to map the overall problem with the Knowledge Profile.

Table 6.4.1.2: Mapping with Knowledge Profile

K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
✓	✓	✓	✓	✓

Justification for EP Attributes Mapping:

- **EP1 - Depth of Knowledge Required:** This research requires high-level expertise in machine learning (deep learning, self-supervised learning with MoCo-v3, Vision transformers like ViT-B/16, TinyViT-5M), biomedical engineering (skin lesion characteristic, dermoscopic images, class imbalance), statistics (custom model validation, accuracy, F1-score) and computational optimization (Multi-Axis Attention, LoRA).
- **EP2 - Range of Conflicting Requirements:** Accuracy vs. Computational Efficiency: Enhanced ViT teacher models promise good accuracy at a cost of extra resources and the TinyViT student model provides a trade-off in resource-constrained environments through knowledge distillation. Generalization vs. Overfitting: Curriculum learning and adaptive weighting address overfitting the models on imbalanced data, leading to reliable melanoma detection.
- **EP3 - Depth of Analysis:** A comprehensive analysis of architectures (ViT-B/16, TinyViT-5M) with metrics such as accuracy, precision, recall and interpretability tools (Grad-CAM) embodies a strong analytical approach for complex problem-solving.
- **EP4 - Familiarity of Issues:** Addresses non-routine challenges such as deeper feature extraction for unlabeled dermatological images, noise in knowledge lesion data, and multi-component loss functions in knowledge distillation go beyond

common conventions.

- **EP5 - Extent of Applicable Codes:** The extent of the Applicable Code is not fulfilled. The project does not include formal adherence to engineering standards or legal codes (ISO, IEEE, medical regulations) rather, it uses best practices of machine learners and software engineers.
- **EP6 - Extent of Stakeholder Involvement:** Involves co-operation between artificial intelligence (model design), biomedical engineering (clinical data preparation) and software engineering (system deployment), with feedback from clinical stakeholders for clinical validation.
- **EP7 - Inter-dependence:** Satisfied by incorporating the integration work of artificial intelligence, biomedical engineering and software engineering expertise, and ensuring good development and deployment of the classification system.

Justification for Knowledge Profile Mapping (linked to EP1):

- **K3 - Engineering Fundamentals:** A solid background in algorithms, data structures, probability and linear algebra is essential to implement self-supervised learning using vision transformer techniques, knowledge distillation processes.
- **K4 - Specialist Knowledge:** Advanced proficiency is required in machine learning (self-supervised learning with MoCo-v3, Multi-Axis Attention, LoRA and curriculum learning), biomedical engineering (dermoscopic image analysis and skin lesion datasets, class imbalance) and statistical modeling.
- **K5 - Engineering Design:** Includes an exercise of creating innovative frameworks by combining self-training, pre-training and improved teacher models (ViT-B/16 with Multi-Axis Attention and LoRA), creating a multi-component loss function for knowledge distillation, overfitting & imbalance management and efficient student models (TinyViT-5M) for resource-limited deployment.
- **K6 - Engineering Practice:** Encompasses rigorous experimental design and statistical validation of the results (using metrics such as accuracy and F1-score), version control for both code and data, implementing tools for interpreting performance such as Grad-CAM, and preparing for clinical application.
- **K8 - Research Literature:** The thesis builds on and contributes to state-of-the-art and cutting-edge research in self-supervised learning, knowledge distillation, Vision Transformers, AI in dermatology and medicine imaging, and therefore requires constant literature review.

6.4.2 Engineering Activities

In this section, we provided a mapping with engineering activities. For each mapping add the rationale by using the Table 6.4.2.

Mapping with Complex Engineering Activities:

Table 6.4.2: Mapping with Complex Engineering Activities

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	×	✓	✓	✓

Justification for Mapping with Complex Engineering Activities:

- **EA1 - Range of Resources:** The project benefits from cutting-edge computing tools (GPUs), dermatological image data and cutting-edge software (PyTorch, MoCo-v3, LoRA), resulting in technical precision and a practical applicability in medical applications.
- **EA2 - Level of Interaction:** Level of Interaction is not met. While there was some collaboration with supervisors and potential clinical experts, there was no significant counterpart in terms of real-time interaction and negotiation of different stakeholder groups throughout the design and testing phase.
- **EA3 - Innovation:** The project is innovative in combining self-supervised learning (MoCo-v3) with Multi-Axis Attention and LoRA in Vision Transformer and optimization of a knowledge distillation process with curriculum learning to provide a novel way for efficient skin lesions classification.
- **EA4 - Consequences for Society and Environment:** This project improves early detection of skin cancer and its impact on patients and public health, and also optimizes the use of computational resources that reduce its impact on the environment while respecting ethical norms (data privacy, fairness, etc.).
- **EA5 - Familiarity:** The project uses innovative but partially familiar approaches (self-supervised learning and knowledge distilling), adapted to the particular challenge of classifying skin lesions, helping to enable reliable and scalable medical diagnostics.

6.5 Summary

This chapter presents the evolution of the proposed knowledge distillation framework for skin cancer classification, combining self-supervised learning, sophisticated model architectures and distillation techniques. It deals with engineering standards, ethics and environmental impacts and focuses upon enhancing the accuracy and efficiency of diagnoses. The project's emphasis on early detection, resource efficiency, and scalability lends itself to possible clinical translation, and adherence to best practices helps preserve the technical and ethical integrity of the data.

Chapter 7

Conclusion

In this chapter, the conclusion of the skin cancer detection research has been presented, with the limitations that we faced and what future work could be done to further increase the capabilities of the system and its effect.

7.1 Summary

In this research, we have developed a novel knowledge distillation framework with the goal of enhancing the efficiency and accuracy of the classification of skin cancer based on dermoscopy images. This work addressed some key challenges, such as class imbalance, limited computational resources and the need for robust feature extraction in different appearances of skin lesions. We combined self-supervised learning (SSL) with multi-axis attention mechanism and knowledge distillation to develop a teacher-student model architecture that emphasizes performance and employability. The teacher model, based on a ViT backbone, enhanced with the Low-Rank Adaptation (LoRA) and multi-axis attention, focused on learning complex lesion patterns using unsupervised pre-training with the help of MoCo-v3, followed by fine-tuning the model on labeled data. The student model, a light-weight TinyViT with specific adaptations, absorbed knowledge from the teacher through a curriculum-based progressive distillation process and achieved similar results using only about 5 million parameters.

We adopted this strategy because skin cancer represents one of the major health issues globally, and current methods have multiple shortcomings that arise due to a lack of data, over-fitting to the majority classes and high computational demand that limits their practical use in real-world applications, especially in resource-constrained clinical environments. As a result, we developed a novel framework that has great potential for early detection of skin cancer through accurate classification in an efficient way that will play a role in saving many lives. By incorporating SSL to achieve better generalization without relying too much on label data, multi-axis attention to focus on features of interest and knowledge distillation with LoRA to reduce the model size, our framework aimed to fill these gaps and make advanced diagnostics more readily accessible and efficient.

To accomplish our goal, we first preprocessed the images using various techniques such as hair removal, resizing and medical-aware augmentations to increase the image's quality and diversity. We then used dual-level class balancing techniques such as oversampling underrepresented classes and focal loss during training to address class imbalances in the datasets. The proposed model was tested on three benchmarking datasets, including HAM10000, ISIC-2019 and PAD-UFES-20 across different metrics such as accuracy, precision, recall, and F1-score. Comparative experiments showed our teacher model outperformed both ensemble and pre-trained models, such as SwinV2-b, ConvNeXt-B, ResNet101 and ViT variants with a significant accuracy improvement even with reduced model size in the student model while maintaining high accuracy and efficient performance. Ablation studies validated the role of LoRA, multi-axis attention, knowledge distillation, and explainable AI techniques such as Grad-CAM and uncovered the model's ability to focus on clinically relevant features. Overall, this research shows one step toward a practical solution for reliable, efficient detection of skin cancer that could aid in its clinical adoption and offer better outcomes for skin cancer patients.

7.2 Limitation

Although advanced techniques such as focal loss and dual-level balancing were used in our framework to handle class imbalance, the models still struggled to accurately classify rare lesion types with a very low number of samples in the datasets. These techniques helped reduce the effect of bias towards majority classes but could not completely mitigate the natural constraint of data scarcity, and led to some misclassification in underrepresented classes. While the overall accuracy was extraordinary, in the real world further refinements are needed to make sure the framework is robust in a variety of clinical settings, with different image quality or hardware constraints. Additionally, our evaluations were restricted to the three popular datasets that may not be fully representative of the diversity of skin variations and new types of lesions globally, which raises the need for more extensive tests in order to increase generalizability.

7.3 Future Work

Future work of this study could explore the integration of lightweight generative models, such as GANs or diffusion-based models for generating high-quality synthetic images of underrepresented classes that could effectively solve the issue of class imbalance and enhance the model's performance across classes. Another key direction is to further optimize the student model using quantization or pruning in order to get real-time inferences for edge devices such as mobile or low-power hardware without dropping accuracy. This study also plans on exploring unsupervised or semi-supervised learning extensions for reducing dependency on large labelled datasets, which could also increase the applicability of the framework for other medical imaging tasks. Finally, making the models more interpretable using other advanced explainable AI techniques, like including integrated attention visualization techniques or a clinician feedback loop, will be key to establishing trust and seamless adoption in healthcare practices.

References

[1] World Cancer Research Fund, "Skin cancer statistics | world cancer research fund," *World Cancer Research Fund*, Oct. 04, 2024. https://www.wcrf.org/preventing-cancer/cancer-statistics/skin-cancer-statistics/

[2] R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal, "Cancer statistics, 2025," *CA: A Cancer Journal for Clinicians*, vol. 75, no. 1, Jan. 2025, doi: https://doi.org/10.3322/caac.21871.

[3] American Cancer Society, "Melanoma Skin Cancer Statistics," *www.cancer.org*, Jan. 12, 2023. https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html

[4] M. Wang, X. Gao, and L. Zhang, "Recent global patterns in skin cancer incidence, mortality, and prevalence," *Chinese medical journal*, vol. 138, no. 2, p. 10.1097/CM9.0000000000003416, 2024, doi: https://doi.org/10.1097/CM9.0000000000003416.

[5] Skin Cancer Foundation, "Skin Cancer Facts & Statistics," *The Skin Cancer Foundation*, Feb. 2024. https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/

[6] WHO, "Skin cancer – IARC," *www.iarc.who.int*, 2022. https://www.iarc.who.int/cancer-type/skin-cancer/

[7] American Academy of Dermatology Association, "Skin Cancer," *American Academy of Dermatology Association*, Apr. 22, 2022. https://www.aad.org/media/stats-skin-cancer

[8] S. Federico, F. Fortarezza, G. Ingravallo, and G. Cazzato, "Epidemiology of Skin Cancer in 2024," *Skin Cancer - Past, Present and Future [Working Title]*, Jan. 2025, doi: https://doi.org/10.5772/intechopen.1008698.

[9] Amdad Hossain Roky *et al.*, "Overview of skin cancer types and prevalence rates across continents," *Cancer Pathogenesis and Therapy*, vol. 3, no. 2, Aug. 2024, doi: https://doi.org/10.1016/j.cpt.2024.08.002.

[10] S. Waseh and J. B. Lee, "Advances in melanoma: epidemiology, diagnosis, and prognosis," *Frontiers in Medicine*, vol. 10, Nov. 2023, doi: https://doi.org/10.3389/fmed.2023.1268479.

[11] R. Kaur, H. GholamHosseini, and M. Lindén, "Advanced Deep Learning Models for Melanoma Diagnosis in Computer-Aided Skin Cancer Detection," Sensors, vol. 25, no. 3, p. 594, Jan. 2025, doi: https://doi.org/10.3390/s25030594.

[12] H. L. Gururaj, N. Manju, A. Nagarjun, N. Manjunath, and F. Flammini, "DeepSkin:

A Deep Learning Approach for Skin Cancer Classification," IEEE Access, vol. 11, pp. 50205–50214, Jan. 2023, doi: https://doi.org/10.1109/access.2023.3274848.

[13] "Skin Cancer Detection Based on Deep Learning," Journal of Biomedical Physics and Engineering, vol. 12, no. 6, Dec. 2022, doi: https://doi.org/10.31661/jbpe.v0i0.2207-1517.

[14] A. Ameri, "A Deep Learning Approach to Skin Cancer Detection in Dermoscopy Images," Journal of Biomedical Physics and Engineering, vol. 10, no. 6, Dec. 2020, doi: https://doi.org/10.31661/jbpe.v0i0.2004-1107.

[15] T. Tuncer, P. D. Barua, I. Tuncer, S. Dogan, and U. R. Acharya, "A lightweight deep convolutional neural network model for skin cancer image classification," Applied Soft Computing, vol. 162, p. 111794, Sep. 2024, doi: https://doi.org/10.1016/j.asoc.2024.111794.

[16] "Skin Cancer ISIC," www.kaggle.com. https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic

[17] H. Vega-Huerta et al., "Classification Model of Skin Cancer Using Convolutional Neural Network," Ingénierie des systèmes d information, vol. 30, no. 02, pp. 387–394, Feb. 2025, doi: https://doi.org/10.18280/isi.300210.

[18] S. Islam, S. T. Alvi, and Era, "Performance Assessment of Advanced CNN and Transformer Architectures in Skin Cancer Detection," pp. 1–8, Sep. 2024, doi: https://doi.org/10.1109/eecsi63442.2024.10776508.

[19] C. Flosdorf, J. Engelker, I. Keller, and N. Mohr, "Skin Cancer Detection utilizing Deep Learning: Classification of Skin Lesion Images using a Vision Transformer," arXiv.org, 2024. https://arxiv.org/abs/2407.18554

[20] X. Zhang et al., "DermViT: Diagnosis-Guided Vision Transformer for Robust and Efficient Skin Lesion Classification," Bioengineering, vol. 12, no. 4, pp. 421–421, Apr. 2025, doi: https://doi.org/10.3390/bioengineering12040421.

[21] C. Xin et al., "An improved transformer network for skin cancer classification," Computers in Biology and Medicine, vol. 149, p. 105939, Oct. 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.105939.

[22] Leema, A. et al. (2025) LMS-VIT: A multi-scale vision transformer approach for real-time smartphone-based skin cancer detection, Frontiers. Available at: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1612502/abstract

[23] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, no. 1, Aug. 2018, doi: https://doi.org/10.1038/sdata.2018.161.

[24] I. Pacal, B. Ozdemir, J. Zeynalov, H. Gasimov, and N. Pacal, "A novel CNN-ViT-based deep learning model for early skin cancer diagnosis," Biomedical Signal Processing and

Control, vol. 104, p. 107627, Jan. 2025, doi: https://doi.org/10.1016/j.bspc.2025.107627.

[25] B. Ozdemir and I. Pacal, "An innovative deep learning framework for skin cancer detection employing ConvNeXtV2 and focal self-attention mechanisms," Results in Engineering, vol. 25, p. 103692, Mar. 2025, doi: https://doi.org/10.1016/j.rineng.2024.103692.

[26] G. Dagnaw^, E. Mouhtadi, and M. Mustapha, "Skin cancer classification using vision transformers and explainable artificial intelligence," J Med Artif Intell, 2024, doi: https://doi.org/10.21037/jmai-24-6.

[27] J. V. Tembhurne, N. Hebbar, H. Y. Patil, and T. Diwan, "Skin cancer detection using ensemble of machine learning and deep learning techniques," Multimedia Tools and Applications, Feb. 2023, doi: https://doi.org/10.1007/s11042-023-14697-3.

[28] N. Codella et al., "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1902.03368 [cs], Mar. 2019, Available: https://arxiv.org/abs/1902.03368

[29] S. Saha, Md. M. Hemal, Md. Z. A. Eidmum, and M. F. Mridha, "Knowledge distillation approach for skin cancer classification on lightweight deep learning model," Healthcare Technology Letters, vol. 12, no. 1, Jan. 2025, doi: https://doi.org/10.1049/htl2.12120.

[30] P. Liu, W. Qian, H. Li, and J. Cao, "A relationship-aware mutual learning method for lightweight skin lesion classification," Digital Communications and Networks, May 2024, doi: https://doi.org/10.1016/j.dcan.2024.04.004.

[31] J. P. Dominguez-Morales, Juan-Carlos Hernández-Rodríguez, L. Duran-Lopez, Julián Conejo-Mir, and Jose-Juan Pereyra-Rodriguez, "Melanoma Breslow Thickness Classification using Ensemble-based Knowledge Distillation with Semi-supervised Convolutional Neural Networks," IEEE Journal of Biomedical and Health Informatics, pp. 1–14, Jan. 2024, doi: https://doi.org/10.1109/jbhi.2024.3465929.

[32] S. Polesie et al., "Can Dermoscopy Be Used to Predict if a Melanoma Is In Situ or Invasive?," Dermatology Practical & Conceptual, p. 2021079, May 2021, doi: https://doi.org/10.5826/dpc.1103a79.

[33] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 538–546, Mar. 2019, doi: https://doi.org/10.1109/jbhi.2018.2824327.

[34] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," IEEE Xplore, Apr. 01, 2018. https://ieeexplore.ieee.org/abstract/document/8363547

[35] Md. A. Rahman, N. M. Fahad, M. A. K. Raiaan, M. Jonkman, F. De Boer, and S. Azam, "Advancing skin cancer detection integrating a novel unsupervised classification and enhanced imaging techniques," CAAI Transactions on Intelligence Technology, Feb. 2025, doi: https://doi.org/10.1049/cit2.12410.

[36] "Skin Lesion Images for Melanoma Classification," www.kaggle.com. https://www.kaggle.com/datasets/andrewmvd/isic-2019

[37] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685 [cs], Oct. 2021, Available: https://arxiv.org/abs/2106.09685

[38] A. R. Khan and A. Khan, "Multi-axis vision transformer for medical image segmentation," *Engineering Applications of Artificial Intelligence*, vol. 158, p. 111251, Jun. 2025, doi: https://doi.org/10.1016/j.engappai.2025.111251.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv:1503.02531 [cs, stat], Mar. 2015, Available: https://arxiv.org/abs/1503.02531

[40] A. G. C. Pacheco et al., "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," Data in Brief, vol. 32, Aug. 2020, doi: https://doi.org/10.1016/j.dib.2020.106221.

Appendix-A

Plagiarism Check

213-15-4583

ORIGINALITY REPORT

19%	12%	13%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Submitted to University of Ulster Student Paper	1%
4	link.springer.com Internet Source	<1%
5	Submitted to University of Sydney Student Paper	<1%
6	Submitted to United International University Student Paper	<1%
7	Submitted to Sreenidhi International School Student Paper	<1%
8	www.ncbi.nlm.nih.gov Internet Source	<1%
9	Submitted to National College of Ireland Student Paper	<1%
10	Submitted to University of Hertfordshire Student Paper	<1%
11	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1%
12	summit.sfu.ca	