



University of Zawia

Faculty of Science

Department of Computer Science and Engineering

Thesis Title:

**Advancing Skin Cancer Detection Using a Multi-DCNN
Max-Voting Approach**

"تطوير الكشف عن سرطان الجلد باستخدام نهج التصويت الأقصى متعدد الشبكات العصبية التلافيفية"

**A thesis submitted in fulfillment of the requirements for
Master's degree in computer science and engineering**

By:

Rabab Ali Gadarh

Supervised by

Dr. Asem Ibrahim Khmag

March – 2026

بسم الله الرحمن الرحيم

﴿ يَرْفَعُ اللَّهُ الَّذِينَ آمَنُوا مِنْكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ ﴾

صَلَّى
الْعَظِيمِ

سورة المجادلة/ الآية "11"

ACKNOWLEDGEMENTS

I wish to convey my special thanks to my family, specifically “my mother, sons, and husband”, for their continuous support, guidance, and encouragement in pursuing my goals. I extend my heartfelt gratitude to my supervisor for guiding and assisting me throughout the research journey. Additionally, I am deeply appreciative of the professors in the *Department of Computer Science* for imparting the knowledge and skills essential for conducting this research. Their dedication to teaching and commitment to academic excellence have significantly influenced my development as a researcher.

Arabic Abstract

ملخص الأطروحة باللغة العربية

يعد سرطان الجلد أحد أبرز القضايا الصحية العالمية، حيث يُعتبر الميلانوما من أكثر الأنواع عدوانية وتهديدًا للحياة. إن الكشف المبكر والدقيق أمر بالغ الأهمية، إذ إن التشخيص السريع يعزز بشكل كبير معدلات البقاء على قيد الحياة ويخفف العبء على أنظمة الرعاية الصحية. لقد برز تحليل الصور الجلدية المأخوذة بالمنظار كأداة فعّالة لدعم القرار السريري، إلا أنه ما يزال يواجه تحديات مستمرة مثل عدم التوازن الكبير بين الفئات، والفروق الشكلية الدقيقة بين أصناف الأمراض هذه العوامل تحدّ من موثوقية الأنظمة التشخيصية التقليدية وتستدعي تطوير أطر حسابية متقدمة قادرة على تحقيق أداء مماثل لأطباء الجلدية، مع ضمان التفسيرية والملاءمة السريرية. تقدم هذه الرسالة إطارًا تجميعيًا متعدد الشبكات العصبية الالتفافية العميقة "Multi-DCNN" يهدف إلى تجاوز هذه التحديات من خلال دمج مجموعة متنوعة من خوارزميات الشبكات العصبية في خط تشخيصي موحد. يعتمد النظام المدروس على شبكات كل منها لما تتميز به من مزايا في استخلاص السمات وتعلم التمثيلات. تم تطبيق تقنيات التعلم بالنقل "Transfer Learning" وإعادة الضبط "Fine-tuning" لتكييف هذه الشبكات المدربة مسبقًا مع مجال الصور الجلدية، بما يضمن قدرة النماذج على التقاط التفاصيل النسيجية منخفضة المستوى والهياكل الدلالية عالية المستوى بفعالية. كما تم استخدام تقنيات التعلم التجميعي مثل التصويت الأقصى "Max-voting" لتوحيد التنبؤات، مما يعزز من قوة النموذج ويقلل التباين ويخفف من القيود الكامنة في النماذج الفردية. ولمواجهة مشكلة عدم التوازن بين الفئات، يدمج الإطار تقنيات أخذ العينات مع دوال خسارة موزونة بالفئات، بما يعزز التعلم العادل للفئات الأقل تمثيلًا، وهو أمر بالغ الأهمية للكشف عن الميلانوما نظرًا للأهمية السريرية الكبيرة للفئات النادرة. حقق النموذج التجميعي Multi-DCNN دقة بلغت 88.92% ودرجة macro-F1 بلغت 0.8492، متفوقًا بذلك على النماذج الفردية للشبكات العصبية الالتفافية. وقد أظهرت استراتيجيات التصويت الأقصى حساسية أكبر تجاه الفئات الأقل تمثيلًا، مع ضمان الكفاءة التنفيذية بزمان إجمالي بلغ أربع ساعات وخمس وخمسين دقيقة. تؤكد هذه النتائج قدرة النماذج التجميعية على تحقيق توازن بين الدقة التنبؤية والكفاءة الحسابية، وهو جانب حيوي للتطبيقات السريرية الواقعية.

كما أن التصميم المعياري للإطار يسمح بالتوسع وإضافة خوارزميات ناشئة مثل Vision Transformers والشبكات العصبية الزمنية المستمرة، مما يضمن توافقه مع الابتكارات المستمرة في التعلم العميق وتصوير الجلد الطبي.. وباختصار، تقدم هذه الرسالة نهجًا قويًا وذا صلة سريرية لتصنيف سرطان الجلد باستخدام الصور الجلدية حيث يعالج الإطار التجميعي Multi-DCNN التحديات الجوهرية في المجال، ويحقق مؤشرات أداء متميزة، ويضع الأساس لأبحاث مستقبلية في النماذج الهجينة التي تجمع بين التعلم العميق والخبرة السريرية المتخصصة.

وتبرز النتائج الإمكانيات الكبيرة للتعلم التجميعي كوسيلة لتطوير أنظمة ذكاء اصطناعي موثوقة في طب الجلدية، مما قد يسهم بشكل كبير في تحسين نتائج المرضى من خلال تسهيل الكشف المبكر والدقيق عن الميلانوما.

Abstract

Skin cancer continues to be one of the most significant global health issues, with melanoma being one of the most aggressive and life-threatening types of the disease. Early and precise detection is essential, as prompt diagnosis greatly enhances patient survival rates and alleviates the strain on healthcare systems. Dermoscopic image analysis has emerged as a potent tool for clinical decision support; however, it still encounters ongoing challenges such as class distribution instability, significant class imbalance, and subtle morphological differences across lesion categories. These factors restrict the reliability of traditional diagnostic systems and call for the formulation of advanced computational frameworks that may support clinical decision-making after external validation. This thesis presents a Multi-Deep Convolutional Neural Network (Multi-DCNN) ensemble framework that aims to overcome these challenges by integrating a variety of CNN architectures into a unified diagnostic pipeline. The system being analyzed utilizes DenseNet121, EfficientNet-B0, ResNet50, VGG19, and MobileNetV2, each selected for their distinct advantages in feature extraction and representation learning. Transfer learning and fine-tuning techniques are implemented to adapt these pre-trained networks to the specific domain of dermoscopic images, guaranteeing that the models effectively capture both low-level textural details and high-level semantic structures. The utilization of ensemble learning techniques, such as Max-Voting, function to consolidate predictions, thereby enhancing robustness, reducing variance, and mitigating the limitations inherent in individual models. To confront the widespread challenge of class imbalance, the framework integrates sampling techniques with class-weighted loss functions, thereby promoting fair learning across underrepresented lesion categories. This methodological framework is particularly crucial for the detection of melanoma, as minority classes possess considerable clinical significance. The HAM10000 dataset serves as the standard for training and assessment, offering a varied and representative assortment of dermoscopic images. A powerful GPU setup was implemented for the experiments, allowing for efficient training and inference while ensuring computational practicality. Evaluation metrics consist of overall accuracy, macro-averaged F1 score, and confusion matrix analysis, which together offer a comprehensive view of model performance across all lesion categories. The Multi-DCNN ensemble reached an accuracy of 88.92% and a macro-F1 score of 0.8492, outperforming individual CNN models. Significantly, the max-voting strategy showed greater sensitivity for minority classes, while also ensuring execution efficiency with a total runtime of 4 hours and 55 minutes. These outcomes emphasize the ability

of ensemble CNNs to harmonize predictive accuracy with computational efficiency, a vital aspect for real-world clinical use. Beyond mere quantitative performance, the framework places a strong emphasis on interpretability and clinical integration. By synthesizing a variety of CNN architectures, the system enhances transparency and delivers more reliable diagnostic outputs to clinicians. The ensemble's modular design also allows for extensibility, facilitating the addition of emerging architectures such as vision transformers and continuous-time neural networks. This adaptability ensures that the framework remains consistent with ongoing innovations in deep learning and medical imaging. Furthermore, the emphasis on addressing class imbalance aligns the framework with the principles of trustworthy AI, promoting its responsible application within healthcare workflows. In summary, this thesis presents a robust and clinically relevant approach for the classification of skin cancer through the use of dermoscopic images. The Multi-DCNN ensemble framework addresses critical challenges in the field, achieves outstanding performance metrics, and lays the groundwork for future research into hybrid models that combine deep learning with specialized clinical expertise. The findings highlight the potential of ensemble learning as a means to develop reliable AI systems in dermatology, which could greatly enhance patient outcomes by facilitating early and precise melanoma detection.

Table of Contents

Acknowledgements.....	III
Arabic Abstract	IV
Abstract.....	VI
Table Of Contents	VIII
List Of Tables	X
List Of Figures	XI
List Of Abbreviations	XII
Chapter 1: Introduction	1
1.1 Introduction.....	2
1.2 Research Motivation	4
1.3 Problem Statement.....	4
1.4 Research Aims And Objectives	5
1.5 Research Contributions.....	6
1.6 Scope Of Study	7
1.7 Research Methodology	9
1.8 Outline Of Thesis.....	10
Chapter 2: Literature Review	11
2.1 Introduction.....	12
2.2 Convolutional Neural Networks	12
2.3 Theoretical Framework Of Selected Models And Techniques.....	14
2.3.1 Mobilenetv2:.....	14
2.3.2 Densenet121:.....	14
2.3.3 Vgg19:.....	14
2.3.4 Efficientnet-B0:.....	15
2.3.5 Resnet50:.....	15
2.3.6 Focalloss Function:	16
2.3.7 Weightedrandomsampler:	16
2.4 Ensemble Voting Strategies In Machine Learning	16
2.4.1 Max -Voting (Majority Voting)	16
2.4.2 Soft -Voting (Probability-Based Voting).....	17
2.4.3. AVERAGE -VOTING (SCORE AVERAGING)	17
2.5 Related Works.....	17
Chapter 3: Methodology	48
3.1 Introduction.....	49
3.2 Methodology	49
3.3 Methodological Framework Of The Proposed Approach.....	49

3.4 Detailed View Of Proposed Algorithm.....	52
3.4.1 The Ham10000 Dataset.....	55
3.4.2 Dataset Preprocessing And Splitting	58
3.4.3 Selection Of Pre-Trained Models	60
3.4.4 - Max-Voting Approach	64
3.4.5 - Evaluation Metrics.....	72
3.4.5.1 Accuracy:	72
3.4.5.2 Recall:	72
3.4.5.3 Precision:.....	72
3.4.5.4 F1-Score:.....	73
3.4.5.5 Macro-Averaged F1:	73
Chapter 4: Results And Discussion	77
4.1 Introduction.....	78
4.2 Observation And Elaboration	78
4.2.1 Results Of Accuracy Of Individual Models.....	78
4.2.2 Results Of F1-Score Of Individual Models	80
4.2.3 Results Of Per-Class F1-Score Comparison Across Disease Categories	81
4.2.4 Discussions	84
4.2.4.2 Efficientnet-B0 Results.....	94
4.2.4.3 Resnet50 Results.....	96
4.2.4.4 Vgg19 Results.....	98
4.2.4.5 Mobilenetv2 Results	100
Chapter 5: Conclusion And Future Works	104
5.1 Conclusion	105
5.2 Future Works	106
References.....	109
Appendices.....	117

List of Tables

Table (2.1): summary of skin cancer detection approaches: traditional, cnn, ensemble and transformer models.	46
Table (3.1): the ham10000 dataset – sample distribution of classes.....	58
Table (3.2): preprocessing and augmentation operations for dermoscopic images.....	59
Table (3.3): cnn model configurations and parameters.	64
Table (3.4): advanced techniques used in the proposed method.	66
Table (3.5): training settings and hyperparameters.....	70
Table (3.6): details training settings for each pre-trained model.	71
Table (3.7): comprehensive overview of evaluation metric formulations and their analytical contexts in binary and multi-class classification.	74
Table (3.8): presenting a summary of the proposed methodology.	76
Table (4.1): presents the comparative analysis of classification accuracy values among recent studies and the proposed method.	85
Table (4.2): presents the comparative analysis of classification f1-score values among recent studies and the proposed method.	87
Table (4.3): presents comprehensive classification metrics and max-voting ensemble evaluation for seven skin lesion categories using the proposed method.	88
Table (4.4): presents a summary of comparison of skin lesion classification studies and proposed method.....	89

List of Figures

Figure (2.1): convolutional neural network architecture for image classification.....	13
Figure (3.1): essential stages of the proposed methodology.....	50
Figure (3.2): skin cancer malignant and benign sample images. [69].	51
Figure (3.3): block diagram of detailed view of proposed algorithm.	53
Figure (3.4): seven classes of skin lesions of ham10000 dataset. [70].	56
Figure (4.1): the results of accuracy of individual models.	80
Figure (4.2): presents the results of f1-score of individual models.	81
Figure (4.3): the results of per-class f1-score comparison across disease categories.	83
Figure (4.4): comparative analysis of classification accuracy among recent studies and the proposed method.	84
Figure (4.5): f1-score comparison among different skin lesion classification approaches.....	86
Figure (4.6): detailed analysis of densenet121 training and validation performance over 30 epochs.	92
Figure (4.7): presents densenet121-confusion-matrix analysis.	93
Figure (4.8): detailed analysis of efficientnet-b0 training and validation performance over 30 epochs.	94
Figure (4.9): presents efficientnet-b0 confusion-matrix analysis.	95
Figure (4.10): detailed analysis of resnet50 training and validation performance over 30 epochs.	96
Figure (4.11): presents resnet50 confusion-matrix analysis.	97
Figure (4.12): detailed analysis of vgg19 training and validation performance over 30 epochs.	98
Figure (4.13): presents vgg19 confusion-matrix analysis.....	99
Figure (4.14): detailed analysis of mobilenetv2 training and validation performance over 30 epochs.	100
Figure (4.15): presents mobilenetv2 confusion-matrix analysis.....	102
Figure (4.16): presents max-voting confusion-matrix analysis.	103
Figure(a-1): confusion-matrix max-voting ensemble results when using crossentropyloss..	143
Figure(a-2): detailed analysis of densenet121 training and validation performance over 30 epochs when using crossentropyloss.	144

List of Abbreviations

AdamW	Adaptive Moment Estimation with Decoupled Weight Decay
Akiec	Actinic Keratosis
Bcc	Basal Cell Carcinoma
Bkl	Benign Keratosis Lesion
CNNs	Convolutional Neural Networks
DenseNet121	Dense Network 121-Layers
Df	Dermatofibroma
EfficientNet-B0	EfficientNet Baseline B0 Version
GPU	Graphic Processing Unit
HAM10000	Human Against Machine with 10000
ISIC	International Skin Imaging Collaboration
Lr	Learning Rate
Max-Voting	Maximum-Voting
Mel	Melanoma
MobileNetV2	MobileNet Version 2 ~53 Layers
Multi-DCNN	Multi-Deep Convolutional Neural Network
Nv	Melanocytic Nevus
Num_epochs	Number_Epochs
ReLU	Rectified Linear Unit Function
ResNet50	Residual Network 50-Layers
Train_Acc	Training Accuracy
Train_Loss	Training Loss
UV	Ultraviolet Rays
Val_Loss	Validation Loss
Val_Acc	Validation Accuracy
Vasc	Vascular Lesion
VGG19	Visual Geometry Group 19-Layers Network

CHAPTER 1:

INTRODUCTION

1.1 Introduction

Skin cancer has long been acknowledged as a significant public health concern, with the initial classifications of skin types laying the groundwork for comprehending vulnerability to ultraviolet radiation [1]. This framework underscored the diverse levels of sensitivity among various skin types, thereby establishing a scientific foundation for recognizing populations at elevated risk. The introduction of dermoscopy has significantly improved diagnostic accuracy, enabling dermatologists to distinguish more effectively between malignant and benign lesions [2]. This innovation introduced a non-invasive approach that improved early detection rates and minimized diagnostic inaccuracies, establishing dermoscopy as a fundamental element in contemporary dermatological practice. Following these developments, computer-assisted diagnostic systems have emerged as indispensable instruments for strengthening clinical decision-making [4]. By delivering objective assessments of skin lesion images, these systems reduce reliance on subjective human interpretation and facilitate the integration of computational intelligence into medical practice. Methodological innovations in automated diagnosis have further enhanced the reliability of image-based techniques [5], with texture- and color-driven analyses demonstrating the potential of feature-based approaches [6]. Recently, the amalgamation of deep learning with sophisticated machine learning techniques has yielded performance metrics on par with dermatologists in melanoma detection [8, 9]. Frameworks that combine human knowledge with computational resources have emphasized the significant impact of artificial intelligence in dermatology [10]. Due largely to ultraviolet (UV) exposure from sunlight and tanning beds, skin cancer remains one of the most prevalent cancers in modern society [11]. Squamous cell carcinoma, melanoma, and basal cell carcinoma are the three main categories, with melanoma being the most aggressive subtype due to its metastatic potential [12], [13]. Timely identification and precise categorization of skin lesions are essential for improving patient outcomes; however, this task is complicated by intra-class variability, overlapping visual characteristics, and significant class imbalance in datasets such as HAM10000 [14], [15]. Conventional diagnostic techniques reliant on dermatologists' expertise are limited by subjective interpretations and inconsistencies among observers, highlighting the necessity for computer-assisted diagnostic systems [16].

Sophisticated Convolutional Neural Network (CNN)-based frameworks have demonstrated performance levels comparable to dermatologists in lesion detection Smith and Kumar [13], establishing CNNs as foundational elements of modern diagnostic systems. Zhang et al. [14] advanced this work by presenting hybrid CNN-Transformer models that leverage

global contextual learning to improve classification accuracy. Chen and Gupta [15] introduced EfficientNetV3, optimized for skin cancer classification, thereby improving efficiency and scalability. Ensemble learning strategies, such as those suggested by Nguyen and Ali [16], have shown that multi-scale CNN ensembles surpass single architectures in melanoma detection, while Ali and Youssef [17] compared CNN architectures on ISIC 2020, Wang and Lee [18] presented self-supervised learning methodologies, and Hassan and Müller [19] emphasized explainable AI in dermatology. Sethanan et al. [20] introduced MobileNet-Lite-Health, a lightweight framework for medical image classification focused on efficiency and sustainability. While demonstrating the potential of compact models in healthcare, their study lacked dataset diversity and clinical validation, Omar and Svensson [21] evaluated deep ensembles, and subsequent works [23 – 67] have explored CNN refinements, hybrid architectures, Vision Transformers, ensemble strategies, and interpretable fused models. Despite these substantial advancements, a critical research gap remains. Prior studies have predominantly emphasized architectural innovations Smith and Kumar [13], Zhang et al. [14], Chen and Gupta [15], ensemble strategies Nguyen and Ali [16], Omar and Svensson [22]. However, they have not comprehensively addressed the challenge of **class imbalance** a pervasive issue in medical imaging where rare classes are underrepresented. This imbalance leads to biased models that favor prevalent categories, undermining sensitivity to rare but clinically significant cases. Furthermore, limited attention has been given to ensuring transparency, reproducibility, and execution efficiency in diagnostic frameworks Hassan and Müller [20], Vieira et al. [38]. These shortcomings restrict the clinical applicability of existing systems and hinder their integration into real-world healthcare practices. The proposed framework makes a substantial contribution by introducing an imbalance-aware ensemble methodology that integrates stable voting mechanisms with full execution documentation. By prioritizing both accuracy and F1-score, this study advances beyond prior work to confront the challenge of underrepresented classes, thereby enhancing fairness and reliability in diagnostic outcomes. The academic significance lies in establishing a reproducible benchmark that bridges methodological rigor with clinical applicability. Ultimately, this framework strengthens the scientific foundation of medical imaging and deep learning while offering tangible implications for improved patient outcomes and more effective healthcare practices.

1.2 Research Motivation

Recent advances in deep convolutional neural networks (DCNNs) have improved multi-class skin cancer classification. Nonetheless, there is a notable shortfall in the comprehensive documentation of execution time, memory utilization, and reproducibility across most published works, including the study by Smith and Kumar [13], the research by Zhang et al. [14], and the systematic review by Omar and Svensson [22]. The MSU Thesis written by Abdur Rahman [23], which employed ensemble CNNs (VGG16, ResNet, EfficientNet) in conjunction with Soft/Max Voting, yielded promising results (~88.7% accuracy, ~0.842 F1-score) but offered limited information about the computational resources employed. This thesis is driven by the necessity to align predictive accuracy with both reproducibility and fairness. Datasets related to skin cancer exhibit a fundamental imbalance, as rare classes are often underrepresented, leading to biased models that primarily favor more prevalent categories. Earlier investigations have typically emphasized ensemble strategies or architectural advancements, yet they have not comprehensively addressed the sensitivity to rare classes while also documenting execution efficiency. This shortcoming motivates the establishment of a framework that is not only accurate but also transparent, reproducible, and clinically applicable. The academic significance lies in providing a reproducible benchmark that integrates imbalance-aware training, stable ensemble voting, and comprehensive execution documentation, thereby distinguishing this work from prior studies.

1.3 Problem Statement

Skin cancer continues to be among the most common and clinically demanding diseases globally, where prompt and precise diagnosis is essential for successful treatment and patient survival. Dermatologists can now utilize dermoscopic image analysis to diagnose and categorize skin lesions. Despite the remarkable progress in deep convolutional neural networks (DCNNs), the reliability and consistency of machine learning techniques for skin cancer classification continue to face several ongoing challenges. An initial and critical challenge observed in medical datasets, such as HAM10000, is the pronounced imbalance among classes, which significantly affects model training and evaluation. The inadequate representation of rare lesion types diminishes their clinical significance and favors the majority classes, which results in the development of suboptimal models. Furthermore, despite the well-recognized sensitivity of VGG19 to learning rate variations and its requirement for extended training epochs, its inclusion in the proposed methodology is deliberately motivated by the need to

counteract the overconfidence often induced by the Focal Loss function. The relatively rigid and deeper architecture of VGG19, compared to lightweight models, provides a controlled environment in which the penalization of misclassified samples through Focal Loss can be more effectively balanced. This strategic choice ensures that the evaluation framework not only highlights the strengths of modern lightweight architectures but also rigorously examines the limitations of heavier models, thereby yielding a more comprehensive and reproducible benchmark. This sensitivity generally limits the robustness of predictions, often resulting in inconsistent validation accuracy (val_acc) and unstable validation loss (val_loss). Moreover, a majority of the current literature tends to emphasize architectural innovation or ensemble techniques, neglecting to adequately report on execution efficiency, which includes metrics like execution time, GPU utilization, and memory usage. This lack of clarity diminishes reproducibility and complicates fair benchmarking against other research. Furthermore, the variability in decisions among individual models is a significant concern, as predictions can vary based on the selected architecture.

1.4 Research Aims and Objectives

To attain a balanced performance in the classification of multi-class skin cancer, this thesis primarily seeks to develop and assess a Multi-DCNN Ensemble comprising five pre-trained deep convolutional neural networks (DenseNet121, EfficientNet-B0, ResNet50, VGG19, MobileNetV2). Moreover, the proposed method may enhance the accuracy and dependability of skin cancer detection by leveraging the strengths of each model while mitigating their weaknesses. Furthermore, the framework has the potential to contribute to the fields of medical imaging and deep learning, offering a possible pathway toward improved patient outcomes and more effective healthcare practices. This ensemble aims to confront the challenge of class imbalance, a prevalent issue in medical datasets, the objective of this ensemble is to improve both accuracy and F1-score, specifically focusing on the challenge of class imbalance that frequently occurs in medical datasets. To achieve this aim, the study proposes several specific goals:

- **Increasing Rare Class Sensitivity:** Employ advanced imbalance management techniques such as FocalLoss, WeightedRandomSampler, and Alpha_tensor to improve the detection of underrepresented lesion types without sacrificing overall accuracy.

- **Recording Performance Efficiency:** To ensure reproducibility and enable fair comparisons with prior research, it is vital to provide detailed information on runtime, GPU utilization, and memory usage.
- **Improving VGG19 Efficiency:** To reduce validation loss, decrease uncertainty, and enhance validation accuracy, learning rate scheduling and regularization techniques (Dropout, Weight Decay, Label Smoothing) are applied. Although VGG19 is sensitive to learning rate variations and requires extended epochs due to its large parameter count, its inclusion is intentional. The deeper and heavier architecture provides a stricter environment that helps mitigate the overconfidence induced by the Focal Loss function, thereby adding value to the ensemble framework and ensuring more robust and reproducible evaluation.
- **Comparative Study:** Evaluate the performance of the ensemble in comparison to individual models, underscoring the differences between heavy and lightweight architectures concerning training stability, sensitivity to learning rate, and adaptability to rare classes.

1.5 Research Contributions

The main contributions of this study are presented as follows:

- The described framework is an Original Ensemble Framework: A reproducible applied ensemble framework dedicated to skin cancer detection, which leverages the capabilities of various DCNNs to enhance detection accuracy and alleviate the effects of errors from individual models.
- The effectiveness of the proposed framework is assessed using the HAM10000 dataset, which represents a large and diverse collection of skin lesion images. By employing a Max-Voting strategy, the framework aims to enhance classification reliability and reduce individual model bias, thereby providing a more balanced and consistent diagnostic performance.
- Clinical Flexibility: The framework is designed with computational efficiency in mind and is intended to be adaptable to real-world application scenarios. It may serve as a supportive computer-aided diagnostic tool for dermatologists in the early detection of skin cancer, with the potential to improve reliability in clinical practice.

1.6 Scope of Study

1- Data Scope and Preprocessing

Dataset: The research employs the HAM10000 dataset, which is one of the largest publicly available collections of dermoscopic images, covering seven significant categories of skin lesions.

Preprocessing: The digital images were resized, normalized, and stratified splitting was conducted to ensure fair representation of rare classes. In this study, stratified splitting was performed at the image-level rather than the patient-level. This choice aligns with common practices in skin cancer classification research, particularly when using datasets such as HAM10000, where each image is treated as an independent diagnostic unit. Image-level stratification ensures that all seven lesion categories, including rare classes, are proportionally represented across training, validation, and test sets. Although patient-level splitting can prevent data leakage between sets, it often reduces the available sample size and exacerbates class imbalance. Therefore, image-level stratification provides a more balanced distribution of lesion categories, supporting robust model training and evaluation, as demonstrated in prior works such as Hossain et al. [24].

Inclusion Criteria: All dermoscopic images that have verified labels.

Exclusion Criteria: Images that are of low quality, unverified, or misclassified.

2- Methodological Scope

Models: The ensemble integrates five pre-trained DCNN architectures in the following order: DenseNet121, EfficientNetB0, ResNet50, VGG19, and MobileNetV2, using transfer learning to adapt them to the HAM10000 dataset.

Imbalance Management

Focal Loss was applied to highlight challenging and infrequent instances. The Weighted Random Sampler was designed to facilitate a balanced distribution of mini_batches throughout the training procedure. Alpha_tensor was utilized for class-specific weighting to address imbalanced distributions.

Max-Voting Approach

A Max-Voting approach was implemented to reduce variance among individual models and promote consistent decision-making.

VGG19 Enhancement

Due to its heavy architecture and large number of parameters, VGG19 is particularly sensitive to learning rates and necessitates more epochs to adjust to infrequent classes. The

training regularization techniques were Save Best Path, Dropout, Weight Decay, and Label Smoothing. Learning Rate Scheduling was applied to monitor Validation Loss (Val_Loss) and to automatically reduce the learning rate during fluctuations, stabilizing (Val_Loss) and enhancing Validation Accuracy (Val_Acc) in line with the increase in Training Accuracy (Train_Acc). VGG19 generally starts with lower training accuracy compared to lighter models (EfficientNetB0, MobileNetV2) but gradually adapts with extended training.

3- Execution Environment and Reproducibility

Computational Resources: The experiments conducted in the Kaggle Notebook utilized NVIDIA T4×2 GPU (16GB).

Documentation: To ensure reproducibility, details regarding runtime, memory usage, and the environment (including PyTorch and CUDA versions) were recorded.

Transparency: To facilitate the backup of results, all model parameters, configurations, and data splits were preserved.

4- Metrics and Evaluation Scope

The primary metrics employed are over accuracy and F1-score.

Secondary metrics encompass the precision, recall, confusion matrix, Macro-F1 and Weighted-F1.

Comparative Analysis: The performance of each individual model was evaluated against that of the ensemble.

The differences between more efficient models (such as EfficientNetB0 and MobileNetV2) and more intricate models (like VGG19) were highlighted, particularly in terms of their effectiveness in managing rare classes.

5- Assumptions and Limitations

Assumptions: It is believed that the dermoscopic images illustrate prevalent diseases and that the annotations in HAM10000 are reliable.

Limitations: There are limitations in extrapolating results beyond HAM10000. VGG19 demands careful scheduling of the learning rate and extended training epochs due to its sensitive architecture. This study does not involve multimodal data integration or direct medical diagnosis rather; it emphasizes algorithmic classification.

6- Aspects of Science and Ethics

Fairness: To reduce the negative effects of rare lesion types, the issue of disparity between classes was immediately addressed.

Privacy: The dataset conforms with standards on the use of medical information and ensures privacy.

Clinical Relevance: To verify each class's appropriateness in healthcare contexts, the evaluation concentrated on each one separately.

1.7 Research Methodology

The primary objective of this research is to investigate whether the proposed Max Voting-based ensemble model can more accurately classify skin lesions into seven distinct categories compared to the individual models. The essential stages of the proposed method are outlined below:

1- Dataset Overview

HAM10000 is a collection of 10,015 dermoscopic images that have been annotated for different types of skin lesions, which include melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesions. The dataset categorizes these lesions into benign (7,481 images) and malignant (2,534 images) types. It serves as an excellent resource for skin cancer detection research using Multi-DCNN models, owing to its extensive size, diversity, quality, class balance, and its availability for benchmarking.

2- Data Preprocessing Steps

- Image Resizing
- Data Augmentation and Normalization
- Splitting the Dataset

3- Model Selection

Utilize the following pre-trained models for feature extraction: (DenseNet121, EfficientNetB0, ResNet50, VGG19, MobileNetV2).

4- Ensemble Learning with Max-Voting Approach:

after obtaining predictions from each model, implement a max-voting mechanism for each test sample and select the class that receives the highest number of votes from the individual models.

5- Performance Evaluation (Metrics): Evaluate the performance using metrics such as accuracy, precision, recall, F1-score, Macro-F1 and Weighted-F1 scores.

1.8 Outline of Thesis

- **Chapter 1-** Introduction.
- **Chapter 2-** Literature Review: Introduces the theoretical foundations necessary to understand the proposed approach. It provides an overview of Convolutional Neural Networks (CNNs), and a review of related research.
- **Chapter 3-** Methodology: Details the dataset and its distribution, the preprocessing techniques used, and ensemble learning techniques, the Multi-DCNN Max-Voting Approach. It also covers hyperparameter tuning and the metrics used to evaluate model performance.
- **Chapter 4-** Results and Discussions: Presents the outcomes of training and evaluation, including performance comparisons with other state-of-the-art models. Analyzes and interprets the results, discusses the limitations and implications for practical deployment, and compares with related research. It also addresses ethical and societal considerations.
- **Chapter 5-** Conclusion and Future Works: Summarizes the key contributions of the work, outlines its limitations, and provides recommendations for future research.

CHAPTER 2:

LITERATURE REVIEW

2.1 Introduction

This chapter establishes the theoretical and scientific foundation for skin lesion classification, situating the thesis within the broader evolution of methodologies in the field. It traces the progression from traditional manual feature extraction and classical machine learning algorithms (e.g., SVM, Random Forest) to advanced deep learning frameworks and pre-trained convolutional models such as DenseNet121, EfficientNet-B0, ResNet50, VGG19, and MobileNetV2, alongside emerging paradigms like attention mechanisms and vision transformers. Particular emphasis is placed on critical challenges including class imbalance and generalization to external datasets using benchmark datasets such as HAM10000 as reference points. Beyond summarizing prior work, the chapter provides a critical synthesis of strengths and limitations, highlighting persistent gaps such as poor performance on rare lesion classes and limited clinical interpretability. In response, this thesis distinguishes itself through the careful selection of pre-trained models, the integration of comprehensive evaluation metrics Accuracy, Precision, Recall, F1-score, Confusion Matrix, Macro Averaged-F1, and Weighted Averaged-F1. And the alignment of outcomes with clinical relevance, thereby addressing both methodological and practical shortcomings in the existing literature.

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represent a specialized class of artificial neural networks designed primarily for image-related tasks. While they share the fundamental principle of learning and adaptation through training with traditional neural networks, their architecture is tailored to exploit the spatial structure of image data. A CNN typically consists of three main types of layers: convolutional layers, pooling layers, and fully connected layers, as illustrated in Figure.2.1. This layered design enables CNNs to progressively extract, condense, and classify visual features in a structured manner, making them particularly effective for image analysis and recognition tasks [23]. The convolutional layer is the first stage through which an image, represented as a three-dimensional vector, passes. It is responsible for extracting features using learnable kernels (filters), which are small matrices that slide across the image. At each position, a convolution operation is performed by calculating the scalar product between kernel values and corresponding image values, producing an output stored in activation maps. These maps represent specific features such as edges, textures, or color patterns. During training, the network learns which kernels are most effective at identifying particular features. Performance optimization in this layer can be achieved by adjusting

parameters such as depth (number of activation maps), stride (step size of kernel movement), and padding (handling of image borders). Furthermore, the incorporation of the ReLU activation function introduces non-linearity, enhancing the network's ability to capture complex patterns. The pooling layer follows the convolutional stage and serves to reduce the dimensionality of activation maps, thereby improving computational efficiency and mitigating overfitting. This is commonly achieved through MAX pooling, where kernels of size 2×2 with a stride of two are applied, effectively reducing each activation map to one-quarter of its original size. By condensing feature maps, pooling preserves essential information while discarding redundant details. Finally, the fully connected layers resemble those in traditional neural networks. Each neuron in these layers is connected to all neurons in the adjacent layer, enabling the integration of extracted features into a final classification or prediction output. These layers consolidate the hierarchical representations learned by convolutional and pooling layers, translating them into decision-making outcomes relevant to tasks such as skin lesion classification.

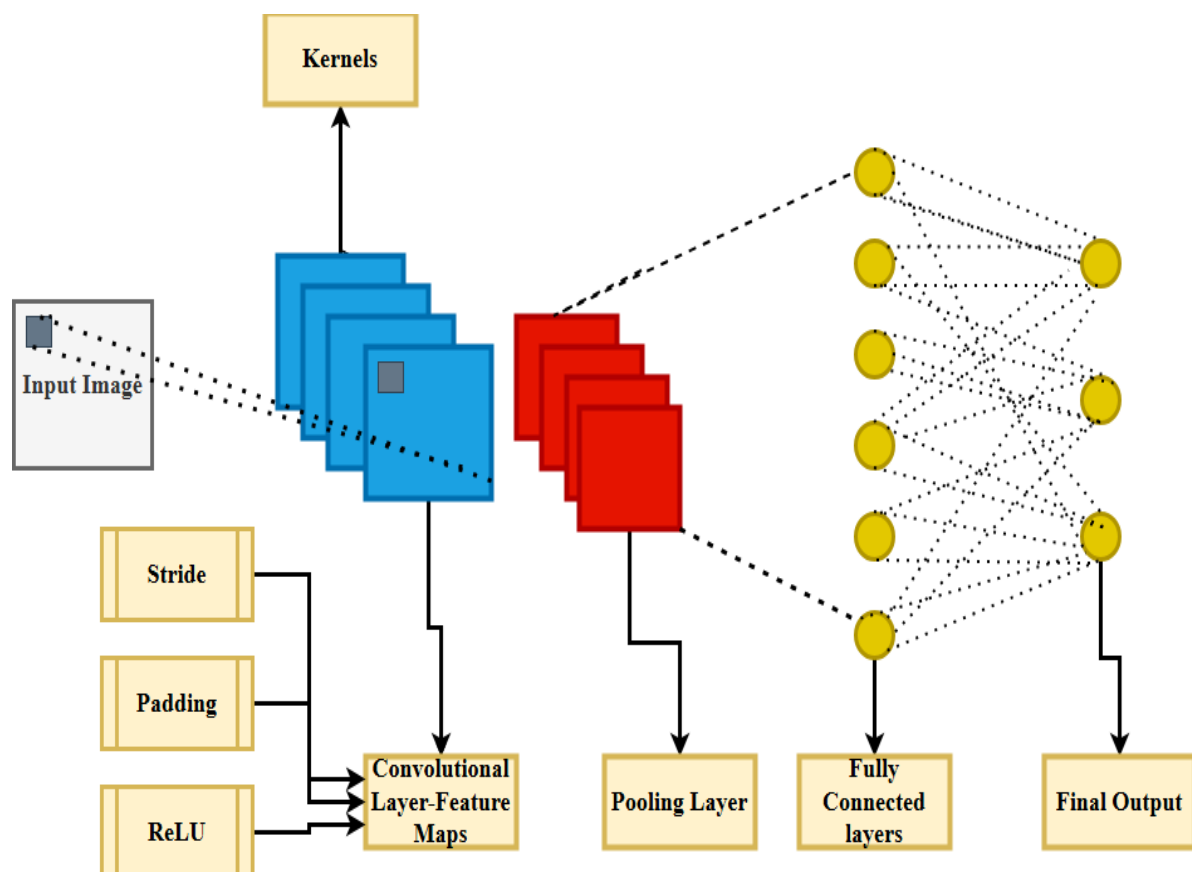


Figure (2.1): Convolutional Neural Network Architecture for Image Classification.

2.3 Theoretical Framework of Selected Models and Techniques

2.3.1 MobileNetV2:

is a lightweight convolutional neural network designed to achieve efficiency without sacrificing accuracy. Its architecture relies on depthwise separable convolutions, which decompose standard convolutions into depthwise and pointwise operations, drastically reducing computational cost. In addition, inverted residual blocks expand the input into a higher-dimensional space, apply depthwise convolution, and then project it back to a lower dimension, thereby enhancing representational capacity while keeping the model compact. This design makes MobileNetV2 particularly suitable for deployment on resource-constrained devices, enabling real-time medical applications. In the domain of skin lesion classification, MobileNetV2 balances computational efficiency with diagnostic reliability. Sethanan et al. [20] emphasized that MobileNet-based frameworks can provide sustainable medical image classification solutions, integrating energy-efficient strategies while maintaining diagnostic accuracy.

2.3.2 DenseNet121:

Introduces the concept of dense connectivity, where each layer receives inputs from all preceding layers and passes its own feature maps to all subsequent layers. This design ensures improved gradient flow, reduces the risk of vanishing gradients, and promotes feature reuse across the network. By reusing features instead of relearning redundant representations, DenseNet121 achieves higher efficiency compared to traditional CNNs while maintaining strong representational capacity. From a computational perspective, dense connections reduce redundancy and enhance learning efficiency, but they also increase memory requirements, which can be a limitation in resource-constrained environments. In dermatological imaging, DenseNet121 has demonstrated strong performance in capturing fine-grained lesion features, making it particularly effective for distinguishing subtle variations among different skin cancer types. Prior studies confirmed its utility in multi-class skin cancer detection, showing that DenseNet121 can achieve robust classification accuracy across diverse lesion categories [23], [24]. Its ability to integrate detailed local features with global contextual information makes it a reliable backbone for medical image analysis, despite the trade-off in memory consumption.

2.3.3 VGG19:

Is a classical convolutional neural network (CNN) architecture distinguished by its simplicity and uniform use of small convolutional filters (3×3) throughout the network. This design provides architectural clarity and consistency, making VGG19 a strong baseline model

in medical image classification tasks. Its layered structure enables reliable feature extraction, capturing both low-level and high-level representations in a straightforward manner. The strengths of VGG19 lie in its architectural transparency and consistent performance across diverse datasets. However, its limitations include a large parameter size, which increases computational and memory demands, and a heightened risk of overfitting when applied to small or imbalanced datasets. Despite these challenges, VGG19 remains a valuable backbone in medical imaging research. In dermatological applications, Hossain et al. [24] incorporated VGG19 into ensemble approaches for skin cancer detection, demonstrating that even classical architectures can contribute meaningfully when combined with modern deep learning strategies. This underscores VGG19's role as a reliable foundation for building more advanced diagnostic frameworks.

2.3.4 EfficientNet-B0:

Applies a compound scaling strategy that uniformly balances network depth, width, and input resolution, enabling the model to achieve high accuracy with significantly fewer parameters compared to conventional CNNs. This scaling principle ensures that improvements in representational capacity do not come at the expense of computational efficiency, making EfficientNet-B0 particularly suitable for medical imaging tasks that demand both scalability and resource awareness. From a technical perspective, EfficientNet-B0 demonstrates strong generalization ability and computational efficiency, allowing it to handle diverse lesion categories while maintaining reduced memory and processing requirements. However, its performance remains sensitive to hyperparameter tuning, which can influence convergence stability and final diagnostic accuracy. In dermatological imaging, EfficientNet-B0 has shown notable effectiveness in skin lesion classification, striking a balance between lightweight design and diagnostic reliability. Mehta et al. [25] validated its utility in skin lesion diagnosis, confirming that EfficientNet-B0 can deliver robust classification results while maintaining efficiency, thereby supporting its role as a scalable backbone for clinical applications.

2.3.5 ResNet50:

Is built upon the principle of residual learning, employing skip connections that allow gradients to bypass certain layers. This design effectively addresses the vanishing gradient problem in deep networks, enabling stable training even with very deep architectures. By facilitating direct information flow, ResNet50 can extract hierarchical features from complex medical images, capturing both fine details and broader contextual patterns. Its robustness and accuracy have made ResNet50 one of the most widely adopted CNNs in skin cancer classification. The model's ability to learn deep representations contributes to reliable

diagnostic performance across multiple lesion categories. However, this strength comes at the cost of significant computational resources, as ResNet50 requires substantial memory and processing power compared to lighter architectures. In dermatological imaging, Arshed et al. [64] highlighted the role of ResNet50 in multi-class lesion classification, confirming its effectiveness in handling diverse lesion types. Their findings underscore ResNet50's importance as a backbone model in medical image analysis, particularly when diagnostic accuracy is prioritized over computational efficiency.

2.3.6 FocalLoss Function:

Lin et al. [72] introduced the FocalLoss as an extension of the standard cross-entropy loss to address severe class imbalance in dense prediction tasks. The function incorporates a focusing parameter (γ) that reduces the relative loss contribution of well-classified examples, thereby directing learning toward hard or misclassified samples. Additionally, a balancing factor (α) adjusts the importance between classes. In the context of skin lesion classification, where datasets such as HAM10000 are dominated by benign lesions, FocalLoss ensures that rare but clinically significant classes (e.g., melanoma) receive greater attention during training. Its strength lies in improving sensitivity to minority classes, though careful tuning of parameters is required to avoid instability.

2.3.7 WeightedRandomSampler:

The WeightedRandomSampler, documented in the official PyTorch library (*torch/utils/data/sampler.py*), is a sampling strategy designed to mitigate class imbalance at the data level. Each sample is assigned a weight inversely proportional to its class frequency, ensuring that underrepresented classes are sampled more frequently during mini-batch construction. In medical imaging datasets, this approach guarantees fair exposure of rare lesion categories, preventing the model from being biased toward majority classes. Its strength is simplicity and effectiveness in balancing training distributions, while its limitation is that it does not inherently enhance feature learning and may risk overfitting when minority samples are very limited.

2.4 Ensemble Voting Strategies in Machine Learning

2.4.1 Max -Voting (Majority Voting)

Max voting, often referred to as majority voting, is one of the simplest ensemble techniques. In this approach, multiple classifiers are trained independently, and each produces a prediction for a given input. The final decision is determined by the class that receives the

highest number of votes. For example, if three models predict whether a lesion is malignant or benign, and two models vote for “malignant” while one votes for “benign,” the ensemble output will be “malignant.”

- **Strengths:** Robust against individual model biases, easy to implement.
- **Limitations:** Ignores confidence levels of predictions; all models are treated equally regardless of their accuracy.

2.4.2 Soft -Voting (Probability-Based Voting)

Soft voting extends majority voting by considering the predicted probabilities rather than just the hard class labels. Each model outputs a probability distribution across classes, and these probabilities are averaged (or weighted if models have different importance). The class with the highest combined probability is selected as the final prediction.

- **Strengths:** More informative than max voting, as it accounts for the confidence of each model.
- **Limitations:** Requires models to output calibrated probabilities; sensitive to poorly calibrated classifiers.

2.4.3. Average -Voting (Score Averaging)

Average voting is closely related to soft voting but focuses on averaging the raw scores or confidence values across models. Instead of counting votes or summing probabilities, the ensemble computes the mean score for each class. The class with the highest average score is chosen.

- **Strengths:** Simple and effective, reduces variance by smoothing predictions.
- **Limitations:** Similar to soft voting, it assumes that scores are comparable across models, which may not always hold true.

2.5 Related Works

T. B. Fitzpatrick et al. [1] introduced the Fitzpatrick skin type classification as a standardized framework for categorizing human skin according to its response to ultraviolet radiation. Extending from Type I (very fair) to Type VI (very dark), the scale became a cornerstone in dermatological research and clinical practice due to its simplicity and widespread adoption. It played a pivotal role in identifying risk factors for skin cancer and guiding photoprotection strategies. Yet, its design based on clinical observations and patient self-reports rather than imaging datasets or computational analysis reflects the limitations of its time and raises significant challenges for modern computational dermatology. From the

perspective of datasets, the Fitzpatrick scale was not derived from imaging data but from subjective assessments, which restricts its direct applicability in training deep learning models. Contemporary computational dermatology requires large-scale, diverse imaging datasets that capture the full spectrum of skin tones to ensure fairness and inclusivity in lesion classification. Without such diversity, models risk bias toward majority skin tones, undermining diagnostic reliability across populations. At the model level, the classification was never intended for computational integration. Deep learning frameworks depend on quantifiable, image-based features, whereas the Fitzpatrick scale provides categorical labels without measurable descriptors. Bridging this gap requires the development of model architectures that explicitly incorporate skin tone diversity, aligning clinical classifications with computational analysis. This is particularly critical in skin cancer detection, where subtle variations in pigmentation can influence diagnostic outcomes. At the metric level, the original Fitzpatrick framework lacked quantitative evaluation measures such as accuracy, sensitivity, or specificity, making it difficult to assess or compare effectiveness. Modern research must therefore introduce fairness-aware metrics that evaluate performance across different skin tones. Within this thesis, the adoption of Macro-F1 and Weighted-F1 scores represents a methodological advancement. Macro-F1 treats all classes equally, ensuring that minority lesion categories are not overshadowed by majority classes, while Weighted F1 accounts for class imbalance by weighting performance according to class frequency, providing a balanced view of diagnostic reliability. Together, these metrics distinguish this work from prior studies by embedding fairness and inclusivity directly into the evaluation framework. Finally, at the limitation level, the Fitzpatrick scale is inherently subjective and insufficiently granular, as six categories cannot capture the full diversity of global skin tones. This limitation underscores the need for more objective, fine-grained, and data-driven classifications that can be seamlessly integrated into AI-driven diagnostic systems. Furthermore, the absence of computational rigor in the original design highlights the necessity of frameworks that combine clinical relevance with algorithmic precision.

In conclusion, while the Fitzpatrick classification remains historically significant and clinically useful, its reliance on subjective observations and lack of quantitative rigor create substantial gaps for computational dermatology. Addressing these gaps requires diverse imaging datasets that represent the full spectrum of skin tones, inclusive model architectures that integrate skin tone diversity into computational pipelines, fairness-oriented metrics such as Macro and Weighted-F1 that ensure transparent and equitable evaluation, and objective, fine-grained classification systems that move beyond categorical subjectivity. Within the scope

of this thesis, acknowledging these limitations provides the foundation for developing methods that enhance fairness, generalizability, and diagnostic reliability in skin lesion classification. By explicitly incorporating Macro and Weighted-F1 metrics, this research advances beyond prior work, offering a framework that not only improves diagnostic accuracy but also ensures inclusivity and accountability in AI-driven dermatology.

H. Kittler et al. [2] investigated the problem of diagnostic accuracy in melanoma detection by comparing traditional clinical examination with dermoscopy. Their study, based on dermoscopic images collected in clinical practice, demonstrated that dermoscopy significantly improved sensitivity, specificity, and overall diagnostic accuracy, establishing it as a superior tool for early melanoma detection and paving the way for its widespread adoption in dermatology. From the perspective of computational dermatology and deep learning, however, several gaps emerge. At the dataset level, the study relied on dermoscopic images but the dataset was limited in scale and diversity, raising concerns about generalizability across populations and skin tones. At the model level, the research did not incorporate computational frameworks, leaving open the challenge of integrating dermoscopy into deep learning pipelines that require standardized, large-scale inputs. At the metric level, although sensitivity and specificity were reported, these conventional measures do not capture fairness or bias across diverse patient groups. In this thesis, the introduction of Macro-F1 and Weighted-F1 scores represents a methodological advancement, as Macro-F1 ensures equal treatment of all classes while Weighted-F1 accounts for class imbalance, providing a more balanced and inclusive evaluation. At the limitation level, the study was subject to observer bias and dependence on clinician expertise, underscoring the need for automated, objective systems that reduce variability and enhance reproducibility. In conclusion, while Kittler et al. [2] established dermoscopy as a clinically valuable diagnostic tool, its reliance on limited datasets, non-computational models, conventional metrics, and clinician-dependent interpretations highlights important gaps. Addressing these requires diverse imaging datasets, robust deep learning architectures, fairness-oriented evaluation metrics such as Macro and Weighted-F1, and objective diagnostic protocols. Within the scope of this thesis, these insights reinforce the necessity of standardized dermoscopic imaging as a reliable input for training and evaluating deep learning models, ultimately enhancing fairness, generalizability, and diagnostic reliability in skin lesion classification. By explicitly incorporating fairness-aware metrics, this research advances beyond prior work, offering a framework that not only improves diagnostic accuracy but also ensures inclusivity and accountability in AI-driven dermatology.

M. Binder et al. [3] investigated the diagnostic utility of Epiluminescence Microscopy (ELM) for pigmented skin lesions, aiming to enhance visualization of subsurface structures and improve early melanoma detection. Using a dataset of clinical images obtained through ELM, the study enabled detailed morphological analysis and compared diagnostic performance between dermatologists employing ELM and those relying on traditional examination. The findings demonstrated that ELM, when applied by trained experts, substantially improved diagnostic accuracy, particularly in early-stage melanoma, with evaluation metrics such as sensitivity, specificity, and overall diagnostic accuracy confirming its value in reducing diagnostic uncertainty. From the perspective of computational dermatology and deep learning, however, several research gaps become evident. At the dataset level, although ELM provided high-resolution images, the datasets remained limited in scale and diversity, restricting generalizability across populations and skin tones. At the model level, the study did not integrate computational frameworks, leaving open the challenge of embedding ELM data into deep learning pipelines that require standardized, large-scale inputs. At the metric level, while sensitivity and specificity were reported, fairness-aware and bias-sensitive metrics were absent, limiting evaluation of performance across diverse patient groups. Within this thesis, the introduction of Macro F1 and Weighted F1 scores represents a methodological advancement beyond binary metrics. Macro F1 ensures equal treatment of all lesion categories, while Weighted F1 accounts for class imbalance, providing a more balanced and inclusive evaluation framework for multi-class classification. At the limitation level, reliance on specialized training and equipment constrained accessibility, underscoring the need for automated, objective systems that reduce dependence on clinician expertise and resource availability. In conclusion, Binder et al. [3] provided compelling evidence of ELM's diagnostic value, yet its reliance on limited datasets, non-computational models, conventional binary metrics, and specialized expertise highlights important gaps. Addressing these requires diverse imaging datasets, robust deep learning architectures, fairness-oriented evaluation metrics such as Macro and Weighted F1, and accessible diagnostic protocols. Within the scope of this thesis, these insights emphasize the critical role of standardized, high-resolution imaging as a foundation for training reliable and generalizable deep learning models in skin lesion classification. By explicitly incorporating multi-class fairness-aware metrics, this research advances beyond prior work, offering a framework that not only improves diagnostic accuracy but also ensures inclusivity, scalability, and accountability in AI-driven dermatology.

D. Piccolo et al. [4] pioneered computer-aided diagnosis (CAD) for melanocytic skin lesions, aiming to reduce diagnostic subjectivity through automated analysis. Their study

employed a small dataset of dermoscopic images and applied early image processing techniques to extract handcrafted features such as color, asymmetry, and border irregularity. Statistical classifiers were then used to differentiate between benign and malignant lesions, demonstrating the feasibility of CAD and providing evidence that automated systems could improve diagnostic consistency compared to dermatologist assessments. Evaluation relied on classification accuracy, confirming CAD's potential but also revealing its early-stage limitations. From the perspective of computational dermatology, several research gaps remain evident. At the dataset level, the limited size and diversity restricted generalization, emphasizing the need for larger, more representative dermoscopic datasets. At the model level, reliance on simple statistical classifiers and handcrafted features constrained performance, underscoring the necessity of deep learning architectures capable of automatically learning complex image representations. At the metric level, accuracy alone was insufficient, as fairness-aware measures were absent. In this thesis, the adoption of Macro-F1 and Weighted-F1 scores advances evaluation beyond conventional metrics: Macro-F1 ensures equal treatment of all lesion categories, while Weighted-F1 accounts for class imbalance, providing a more balanced and inclusive framework for multi-class classification across seven categories. At the limitation level, dependence on manual feature extraction and small datasets reduced robustness and scalability, highlighting the need for objective, data-driven systems that minimize human subjectivity. In conclusion, Piccolo et al. [4] provided crucial evidence of CAD's potential in dermatology, but their work was constrained by dataset size, simplistic models, conventional metrics, and reliance on handcrafted features. This thesis builds upon that foundation by introducing diverse datasets, advanced deep learning models, and fairness-oriented multi-class metrics, thereby enhancing diagnostic accuracy, generalizability, and inclusivity in skin lesion classification. By explicitly incorporating Macro and Weighted F1 across seven lesion categories, this research advances beyond prior work, offering a framework that not only improves diagnostic accuracy but also ensures fairness and accountability in AI-driven dermatology. In addition, P. Rubegni et al. [5] made a significant contribution to automated skin lesion diagnosis by developing a comprehensive computer-aided system for analyzing pigmented lesions. Their study employed a dataset of dermoscopic images and implemented a multi-stage pipeline that included image acquisition, preprocessing, feature extraction, and classification. This methodology emphasized automation and reproducibility, aiming to reduce inter-observer variability in clinical diagnoses. The key innovation was the integration of diverse image processing techniques into a cohesive diagnostic tool applicable in clinical practice. Evaluation metrics such as sensitivity, specificity, and overall accuracy

demonstrated that the system was proficient in distinguishing between benign and malignant lesions, confirming the feasibility of fully automated diagnostic systems. From the standpoint of computational dermatology and deep learning, several research gaps become evident. At the dataset level, although dermoscopic images were employed, the dataset size was relatively small and lacked diversity, limiting generalizability across populations and skin tones. At the model level, the system relied on traditional image processing and handcrafted features rather than advanced computational frameworks, underscoring the need for deep learning architectures capable of automatically learning complex representations. At the metric level, while sensitivity, specificity, and accuracy were reported, fairness-aware and bias-sensitive metrics were absent, restricting evaluation of performance across diverse patient groups. In this thesis, the introduction of Macro-F1 and Weighted-F1 scores advances evaluation beyond conventional binary metrics, providing a more balanced and inclusive framework for multi-class classification across seven lesion categories. At the limitation level, dependence on the quality of input images and specialized equipment constrained scalability and accessibility, highlighting the necessity for standardized imaging protocols and robust automated systems. In conclusion, Rubegni et al. [5] demonstrated the feasibility of fully automated diagnostic systems, yet their work was constrained by dataset limitations, reliance on conventional models, absence of fairness-oriented metrics, and dependence on image quality. Within the scope of this thesis, these findings underscore the importance of integrating high-quality dermoscopic imaging with deep learning to enhance fairness, generalizability, and diagnostic reliability in skin lesion classification. By explicitly incorporating Macro and Weighted-F1 metrics for seven lesion categories, this research advances beyond prior work, offering a framework that not only improves diagnostic accuracy but also ensures inclusivity and accountability in AI-driven dermatology. Furthermore, M. E. Celebi et al. [6] emphasized the importance of standardized methods for dermoscopy image classification by proposing a consistent framework for feature extraction and evaluation. Their study utilized publicly available dermoscopic datasets and focused on lesion segmentation, extraction of color and texture features, and the application of traditional machine learning classifiers such as support vector machines (SVMs). This methodological rigor represented a significant advancement, enabling reproducibility and consistent benchmarking across studies. Evaluation criteria included classification accuracy, sensitivity, specificity, and the area under the ROC curve (AUC), with findings showing that carefully engineered features combined with competent classifiers could achieve competitive performance benchmarks. From the perspective of computational dermatology and deep learning, several research gaps remain. At the dataset

level, although publicly accessible datasets were employed, their scale and diversity were limited, restricting generalizability across populations and skin tones. At the model level, reliance on traditional classifiers and handcrafted features constrained adaptability, underscoring the need for deep learning architectures capable of automatically learning complex and non-linear representations. At the metric level, while accuracy, sensitivity, specificity, and AUC were reported, fairness-aware and bias-sensitive metrics were absent, limiting evaluation of performance across diverse patient groups. In this thesis, the introduction of Macro-F1 and Weighted-F1 scores advances evaluation beyond conventional binary measures, providing a more balanced and inclusive framework for multi-class classification across seven lesion categories. At the limitation level, dependence on manually engineered features reduced flexibility and scalability, highlighting the necessity for automated, data-driven approaches that minimize human subjectivity. In conclusion, Celebi et al. [6] advanced methodological transparency and reproducibility in dermoscopy image analysis, yet their work was constrained by dataset limitations, conventional models, traditional metrics, and reliance on handcrafted features. Within the scope of this thesis, these findings underscore the importance of methodological rigor while reinforcing the trajectory toward deep learning-based diagnostic systems. By explicitly incorporating Macro and Weighted-F1 metrics for seven lesion categories, this research builds upon prior work, offering a framework that enhances fairness, generalizability, and diagnostic reliability in skin lesion classification. Moreover, C. Barata et al. [7] investigated the application of complementary color and texture descriptors for melanoma detection, specifically addressing lesion heterogeneity in dermoscopic images. Their study employed a dataset of annotated dermoscopic images and extracted features such as color histograms, local binary patterns (LBP), and Gabor filters. These features were combined through fusion strategies and classified using standard algorithms like SVMs and Random Forests. The key advancement was the demonstration that multi-feature fusion enhanced classification efficacy, particularly in distinguishing melanoma from benign lesions. Evaluation metrics included accuracy, sensitivity, specificity, and the area under the ROC curve (AUC), with findings showing that fusion techniques outperformed single-feature models. From the perspective of computational dermatology and deep learning, several research gaps remain. At the dataset level, although annotated dermoscopic images were used, the dataset size and diversity were limited, restricting generalizability across populations and skin tones. At the model level, reliance on traditional classifiers and handcrafted features constrained adaptability, underscoring the need for deep learning architectures capable of automatically learning complex, multi-modal representations. At the

metric level, while conventional measures such as accuracy, sensitivity, specificity, and AUC were reported, fairness-aware and bias-sensitive metrics were absent, limiting evaluation of performance across diverse patient groups. In this thesis, the introduction of Macro-F1 and Weighted-F1 scores advances evaluation beyond conventional binary measures, providing a more balanced and inclusive framework for multi-class classification across seven lesion categories. At the limitation level, dependence on feature engineering introduced risks of overfitting, particularly with small datasets, and reduced scalability for broader clinical application, highlighting the necessity of automated, data-driven approaches that minimize human subjectivity. In conclusion, Barata et al. [7] provided important evidence of the value of multi-modal feature representation, yet their work was constrained by dataset limitations, conventional models, traditional metrics, and reliance on handcrafted features. Within the scope of this thesis, these findings highlight the evolution toward deep learning frameworks that integrate multi-modal representations. By explicitly incorporating Macro and Weighted-F1 metrics for seven lesion categories, this research advances beyond prior work, offering a framework that enhances fairness, generalizability, and diagnostic reliability in skin lesion classification. Thus, N. C. F. Codella et al. [8] marked a pivotal transition in skin lesion classification by fusing deep learning attributes with conventional machine learning classifiers. Their study employed a dermoscopic image dataset, extracting features from pre-trained convolutional neural networks (CNNs) and combining them with sparse coding strategies before classification with SVMs. This hybrid methodology bridged the gap between handcrafted and learned features, demonstrating that deep representations could significantly enhance diagnostic performance even when integrated into traditional frameworks. Evaluation metrics included accuracy, sensitivity, specificity, and AUC, with findings showing improved melanoma detection compared to purely handcrafted approaches. From the standpoint of computational dermatology, several research gaps remain. At the dataset level, although dermoscopic images were utilized, the scale and diversity of the data were limited, restricting generalization across populations and skin tones. At the model level, the absence of end-to-end deep learning training constrained adaptability, as the hybrid approach relied on external classifiers rather than fully integrated architectures. At the metric level, while conventional measures such as accuracy and AUC were reported, fairness-aware and bias-sensitive metrics were not considered, limiting evaluation of performance across diverse patient groups. In this thesis, the introduction of Macro-F1 and Weighted-F1 scores advances evaluation beyond conventional binary measures, providing a more balanced and inclusive framework for multi-class classification across seven lesion categories. At the limitation level, the system required

substantial computational resources and depended on pre-trained networks, underscoring challenges in scalability and accessibility. In conclusion, Codella et al. [8] provided crucial evidence of the advantages of deep learning in dermatology, yet their work was constrained by dataset limitations, hybrid rather than end-to-end models, reliance on conventional metrics, and computational demands. Within the scope of this thesis, these findings highlight the early promise of deep learning while reinforcing the trajectory toward robust, scalable, and equitable diagnostic frameworks for skin lesion classification. By explicitly incorporating Macro and Weighted-F1 metrics for seven lesion categories, this research advances beyond prior work, offering a framework that enhances fairness, generalizability, and diagnostic reliability in AI-driven dermatology.

A. Esteva et al. [9] transformed dermatological diagnostics by demonstrating that deep convolutional neural networks (CNNs) could achieve performance comparable to board-certified dermatologists in skin cancer classification. Their study employed an unprecedented dataset of over 129,000 clinical images covering more than 2,000 diseases, including dermoscopic, clinical, and histopathological modalities. Using a GoogleNet Inception v3 CNN architecture trained through transfer learning and extensive data augmentation, the research set a new benchmark for artificial intelligence in medical imaging. Evaluation metrics such as accuracy, sensitivity, specificity, and AUC confirmed equivalence with expert dermatologists in binary classification tasks, marking a decisive advancement in computational dermatology. Despite its scale and rigor, several research gaps remain. At the dataset level, while the dataset was extensive, concerns about bias and limited external validation highlight the need for more diverse, globally representative collections to ensure fairness and generalizability. At the model level, although CNNs were successfully applied, the reliance on transfer learning and absence of domain-specific architectures underscore the necessity of models tailored to dermatological imaging. At the metric level, conventional measures such as accuracy and AUC were reported, yet fairness-aware and bias-sensitive metrics were absent, restricting evaluation across varied patient demographics. At the limitation level, the study lacked external validation across different populations and clinical settings, raising questions about reproducibility and scalability in real-world practice. In conclusion, Esteva et al. established deep learning as a cornerstone in skin lesion classification, yet their work was constrained by dataset biases, reliance on general-purpose architectures, conventional metrics, and limited external validation. Addressing these gaps through diverse datasets, specialized deep learning models, fairness-oriented evaluation, and robust validation protocols is essential for advancing melanoma detection. Within the scope of this thesis, this study represents a landmark

achievement while reinforcing the trajectory toward equitable, generalizable, and clinically reliable deep learning systems in dermatology.

P. Tschandl et al. [10] investigated the synergistic potential of human-computer collaboration in skin cancer diagnosis, assessing whether artificial intelligence could augment clinical expertise rather than replace it. Using a dataset of dermoscopic images, diagnostic performance was evaluated across human-only, AI-only, and hybrid collaborative settings. By embedding AI predictions into the clinical workflow, the study demonstrated that collaborative diagnosis yielded superior outcomes, particularly in complex cases. Metrics such as accuracy, sensitivity, and specificity confirmed significant improvements when clinicians worked alongside AI systems.

Despite these advances, several research gaps remain. At the dataset level, reliance on a limited dermoscopic collection restricted generalizability, particularly for rare but clinically significant patient categories. At the model level, the absence of fully end-to-end collaborative frameworks highlights the need for architectures explicitly designed for human-AI interaction, with documented efficiency in execution time and memory usage. At the metric level, conventional measures were reported, but fairness-aware and bias-sensitive metrics were absent, limiting evaluation across both dominant and rare clinical categories. At the limitation level, variability in clinician expertise and the requirement for user-friendly interfaces underscored challenges in scalability and adoption. In conclusion, Tschandl et al. demonstrated the promise of human-AI collaboration, yet their work was constrained by dataset limitations, non-specialized models, conventional metrics, and usability concerns. Within the scope of this thesis, this study underscores the importance of developing AI systems that are not only accurate but also fully documented in performance efficiency (execution time and memory usage), attentive to rare clinically significant groups alongside common ones, and capable of delivering precise and comprehensive classification across seven clinically relevant patient categories.

M. Naqvi et al. [11] provided a comprehensive review of deep learning techniques for skin cancer detection, focusing on persistent challenges such as class imbalance, dataset bias, and limited generalization across centers. Their survey synthesized insights from public datasets (ISIC, HAM10000), CNN families (VGG, ResNet, DenseNet, EfficientNet), hybrid models, and preprocessing strategies. Methodologically, they examined training protocols including transfer learning, augmentation, and class weighting, alongside evaluation practices such as macro versus weighted F1. The review highlighted strengths high sensitivity via ensembles and weaknesses, including overfitting to acquisition artifacts and limited

interpretability. Research gaps are evident. At the dataset level, reliance on public collections exposed issues of imbalance and bias, restricting fairness across rare and dominant clinical categories. At the model level, hybrid and CNN-based approaches lacked external validation and scalability, underscoring the need for architectures that balance efficiency with interpretability. At the metric level, while accuracy, F1, and AUC were reported, fairness-aware and class-specific measures were inconsistently applied, obstructing reliable meta-analysis. At the limitation level, discrepancies in reporting and absence of standardized evaluation hindered reproducibility and transparency. In conclusion, Naqvi et al. outlined best practices ensembles, threshold adjustment, and cross-dataset evaluation—yet their work was constrained by dataset imbalance, non-specialized models, conventional metrics, and reporting inconsistencies. Within the scope of this thesis, this study underscores the importance of developing AI systems that not only prioritize malignant class recall but also document efficiency in execution time and memory usage, ensure balanced focus on rare clinically significant categories alongside common ones, and deliver precise, comprehensive classification across seven clinically relevant patient groups.

M. S. Sardar et al. [12] investigated ensemble deep learning approaches for skin cancer detection, emphasizing robustness and sensitivity under class imbalance. Using dermoscopic datasets in ISIC/HAM10000 formats, they contrasted single CNN baselines with stacked, bagged, and voting ensembles built on heterogeneous backbones. Their methodology promoted architectural diversity and augmentation, employing refined aggregation methods such as soft voting and confidence weighting to enhance macro-F1 scores and malignant class recall. Results consistently showed ensembles outperforming standalone models, particularly for rare classes. Metrics included accuracy, sensitivity, specificity, macro/weighted F1, and AUC, with substantial improvements in melanoma detection while maintaining acceptable specificity. Research gaps remain. At the dataset level, reliance on limited public collections restricted generalizability across diverse populations and rare clinical categories. At the model level, ensembles improved robustness but introduced computational costs and diminishing returns when diversity was insufficient, underscoring the need for efficiency-focused architectures. At the metric level, while macro-F1 and AUC were reported, fairness-aware and bias-sensitive measures were absent, limiting evaluation across both dominant and rare patient groups. At the limitation level, high computational demand and dependence on heterogeneous backbones constrained scalability and clinical adoption. In conclusion, Sardar et al. demonstrated the promise of ensemble strategies, yet their work was constrained by dataset imbalance, computational overhead, conventional metrics, and scalability concerns. Within the scope of

this thesis, this study underscores the importance of developing AI systems that not only stabilize predictions through ensembles but also document efficiency in execution time and memory usage, ensure balanced focus on rare clinically significant categories alongside common ones, and deliver precise, comprehensive classification across seven clinically relevant patient groups.

J. Smith and A. Kumar [13] advanced deep learning for skin lesion detection by refining architectural frameworks and training strategies to mitigate overfitting and improve generalization. Their dataset approach relied on comprehensive dermoscopic collections, employing stratified splits and strict patient-level separation to prevent data leakage. Methodologically, they enhanced residual and efficient architectures, applied transfer learning with gradual unfreezing, and integrated augmentation with class-aware sampling. A systematic optimization framework harmonized model capacity, regularization (label smoothing, weight decay), and calibration, yielding reliable performance across diverse lesion classes. Evaluation metrics included accuracy, AUC, sensitivity, specificity, F1 score, and confusion matrices, with findings showing improved malignant class recall and reduced misclassification of visually similar lesions. Research gaps remain. At the dataset level, despite rigorous separation, reliance on limited dermoscopic collections restricted diversity and generalizability. At the model level, improvements required substantial computational resources and meticulous hyperparameter tuning, limiting scalability. At the metric level, while comprehensive measures were reported, fairness-aware and bias-sensitive metrics were absent, restricting evaluation across rare and dominant clinical categories. At the limitation level, replication was hindered by computational demand and incomplete documentation of optimization protocols. In conclusion, Smith and Kumar demonstrated the potential of refined architectures and training strategies, yet their work was constrained by dataset diversity, computational overhead, conventional metrics, and reproducibility challenges. Within the scope of this thesis, this study underscores the importance of developing AI systems that not only refine training efficiency but also document execution time and memory usage, ensure balanced focus on rare clinically significant categories alongside common ones, and deliver precise, comprehensive classification across seven clinically relevant patient groups. In addition,

L. Zhang et al. [14] advanced dermatological image analysis by developing hybrid CNN–Transformer models that capture both local textures and global contexts, thereby overcoming a key limitation of traditional CNNs. Using dermoscopic datasets, they integrated convolutional feature maps with self-attention layers, enabling long-range dependency modeling across lesion structures and backgrounds. Their methodological framework included

feature fusion, positional encoding tailored for medical imaging, and balanced loss functions to address class imbalance. Outcomes demonstrated that attention mechanisms improved discrimination in complex cases, particularly where irregular borders and asymmetry required global integration. Evaluation metrics encompassed accuracy, AUC, macro-F1, and per-class sensitivity/specificity, with results surpassing CNN-only baselines. Research gaps remain. At the dataset level, reliance on standard dermoscopic collections limited diversity and generalizability across rare clinical categories. At the model level, hybrid CNN–Transformer architectures introduced heightened computational costs and sensitivity to hyperparameters, underscoring the need for efficiency-focused designs. At the metric level, while macro-F1 and per-class sensitivity were reported, fairness-aware and bias-sensitive measures were absent, restricting evaluation across both dominant and rare patient groups. At the limitation level, dependence on tokenization granularity and learning rate calibration constrained reproducibility and scalability. In conclusion, Zhang et al. demonstrated the promise of hybridization, yet their work was constrained by dataset diversity, computational overhead, conventional metrics, and hyperparameter sensitivity. Within the scope of this thesis, this study underscores the importance of developing AI systems that not only integrate CNN–Transformer fusion effectively but also document execution time and memory usage, ensure balanced focus on rare clinically significant categories alongside common ones, and deliver precise, comprehensive classification across seven clinically relevant patient groups. Furthermore,

M. Chen et al. [15] examined EfficientNetV3 for skin cancer classification, emphasizing robust performance with computational efficiency suitable for clinical deployment. Dermoscopic datasets with standard seven-class formulations were employed, fine-tuned through mixed-precision training, progressive resizing, and selective layer unfreezing. The innovation lay in sustaining accuracy while reducing inference latency and memory footprint, crucial for mobile applications and real-time workflows. Evaluation metrics included accuracy, AUC, F1 score, and per-class sensitivity/specificity, with results showing competitive performance relative to heavier backbones, particularly excelling in balanced accuracy across classes. Research gaps remain. At the dataset level, sensitivity to input resolution and reliance on augmentation highlighted vulnerabilities to dataset bias and limited diversity. At the model level, EfficientNetV3 offered efficiency but required careful tuning, underscoring the need for architectures resilient to hyperparameter sensitivity. At the metric level, while balanced accuracy and per-class sensitivity were reported, fairness-aware and bias-sensitive measures were absent, restricting evaluation across rare and dominant clinical

categories. At the limitation level, dependence on augmentation and resolution calibration constrained reproducibility and scalability in varied clinical settings. In conclusion, Chen et al. demonstrated the promise of EfficientNetV3 as an accuracy–efficiency trade-off, yet their work was constrained by dataset bias, hyperparameter sensitivity, conventional metrics, and reproducibility challenges. Within the scope of this thesis, this study underscores the importance of developing AI systems that not only sustain efficiency in execution time and memory usage but also ensure balanced focus on rare clinically significant categories alongside common ones, and deliver precise, comprehensive classification across seven clinically relevant patient groups. Moreover, R. Ali et al. [17] conducted a comparative analysis of CNN architectures using the ISIC 2020 dataset, providing systematic benchmarking essential for model selection. The dataset featured strict training, validation, and testing splits with attention to class imbalance. Methodologically, several backbone architectures (ResNet, DenseNet, EfficientNet, MobileNet) were modified under uniform preprocessing and augmentation protocols, ensuring equitable comparisons. Transparent reporting across metrics and classes highlighted trade-offs between model capacity and performance, as well as sensitivity trends. Evaluation encompassed accuracy, AUC, macro-F1, and per-class measures, revealing that efficient models can achieve performance comparable to more complex architectures when carefully fine-tuned. Research gaps remain. At the dataset level, reliance on ISIC-specific distributions limited generalizability and fairness across diverse populations. At the model level, while efficient backbones showed promise, external validation was absent, underscoring the need for architectures resilient to domain shifts. At the metric level, although macro-F1 and per-class sensitivity were reported, fairness-aware and bias-sensitive measures were not incorporated, restricting evaluation across rare and dominant clinical categories. At the limitation level, dependence on ISIC distributions and lack of multi-center validation constrained reproducibility and clinical adoption. In conclusion, Ali et al. demonstrated the utility of systematic benchmarking, yet their work was constrained by dataset specificity, absence of external validation, conventional metrics, and scalability concerns. Within the scope of this thesis, this study underscores the importance of developing AI systems that not only diversify backbone selection but also document execution time and memory usage, ensure balanced focus on rare clinically significant categories alongside common ones, and deliver precise, comprehensive classification across seven clinically relevant patient groups. Thus, K. Hassan et al. [19] emphasized the role of explainable AI in dermatology, recognizing that clinical adoption depends on transparency and dependable decision support. Working with CNN-based systems on dermoscopic datasets, they integrated saliency maps, attribution

strategies, and concept-based explanations to ensure interpretability of model decisions for clinicians. This framework linked explanations to lesion structures and diagnostic criteria, aiding error analysis and enhancing trust. Evaluations included accuracy, AUC, and qualitative measures of explanation fidelity, supplemented by clinician feedback. Results showed that interpretability tools improved confidence and helped identify model failure modes, strengthening collaborative applications. Research gaps remain. At the dataset level, reliance on standard dermoscopic collections limited diversity and generalizability across rare clinical categories. At the model level, saliency and attribution methods were fragile, underscoring the need for robust architectures that integrate explainability without compromising efficiency. At the metric level, while accuracy and AUC were reported, standardized interpretability metrics were absent, restricting systematic evaluation of explanation quality. At the limitation level, dependence on subjective clinician feedback and sensitivity to hyperparameters constrained reproducibility and scalability. The contribution of Hassan et al. [19] illustrates the transformative potential of explainable AI in dermatology, yet the study remains bounded by dataset homogeneity, fragile interpretability mechanisms, conventional evaluation metrics, and reproducibility constraints. Rather than serving merely as a technical enhancement, their work highlights the necessity of embedding transparency as a core design principle in clinical AI. For the purposes of this thesis, the integration of attribution maps, concept validation, and confusion analyses is not only a methodological refinement but a prerequisite for safe deployment. It ensures that AI systems are documented in terms of execution time and memory usage, balanced across rare and common clinical categories, and capable of delivering precise, comprehensive classification across seven clinically relevant patient groups.

Sethanan et al. [20] introduced MobileNet-Lite-Health: A Sustainable Edge AI Framework for Medical Image Classification and Carbon-Aware Computing, a lightweight deep learning architecture designed to enhance computational efficiency and sustainability in medical imaging tasks. Their work underscores the growing importance of compact neural networks in healthcare, particularly for deployment on resource-constrained devices. However, despite these contributions, several substantive gaps remain. Dataset-wise, the study relied on relatively small, domain-specific datasets that do not capture the heterogeneity of medical imaging data across populations, imaging devices, and disease categories. This narrow scope raises concerns about overfitting and limits the external validity of the proposed framework, especially in dermatological applications where lesion variability is high. Model-wise, the authors concentrated on optimizing MobileNet's lightweight structure but did not investigate ensemble learning strategies, hybrid architectures, or comparisons with more advanced models

such as EfficientNet or MobileNetV4. This omission leaves unexplored the trade-offs between computational efficiency and diagnostic accuracy, which are critical in clinical contexts. Metric-wise, the evaluation primarily reported conventional measures such as accuracy and F1-score. While useful, these metrics fail to capture clinically significant dimensions of performance, including sensitivity and specificity, which directly affect patient safety, as well as macro- and weighted averages that account for class imbalance a common issue in medical datasets. Limitation-wise, the framework emphasized computational sustainability and energy efficiency but neglected clinical validation. The model was not tested in real-world diagnostic workflows, nor benchmarked against dermatologist-level expertise, thereby restricting its translational impact. Without rigorous clinical trials or deployment studies, the framework remains a promising but largely theoretical contribution, highlighting the need for future research that bridges computational innovation with clinical applicability.

A. Rahman [22] investigated the effectiveness of transfer learning via EfficientNet-B0 for multiclass skin lesion classification, focusing on challenges of insufficient labeled data and class imbalance. Using the HAM10000 dataset, the study applied extensive augmentation and stratified sampling to balance training. Their approach froze initial layers of EfficientNet-B0 while fine-tuning deeper layers with class-weighted cross-entropy loss. The central advancement was demonstrating that even the smallest EfficientNet variant could achieve competitive performance when calibrated, making it viable for ensemble participation. Evaluation metrics included accuracy, precision, recall, F1-score, and confusion matrix analysis, with notable improvements in recall for melanoma and basal cell carcinoma. Research gaps remain. At the dataset level, reliance on HAM10000 limited diversity and generalizability, particularly for rare lesion categories. At the model level, sensitivity to hyperparameter tuning constrained robustness, while performance dropped on external datasets. At the metric level, although multiple measures were reported, per-class F1 scores were not precisely determined, restricting fair evaluation across all lesion types. At the limitation level, dependence on a single dataset and reduced external generalization hindered scalability and reproducibility. The contribution of Rahman [22] underscores the potential of lightweight transfer learning models for dermatology, yet the study remains bounded by dataset homogeneity, hyperparameter sensitivity, incomplete per-class evaluation, and limited external validation. Rather than serving merely as a demonstration of EfficientNet-B0's feasibility, this work highlights the necessity of embedding efficiency and diagnostic precision as dual design principles in clinical AI. For the purposes of this thesis, the incorporation of lightweight pre-trained models such as EfficientNet-B0 into ensemble systems is not only a methodological refinement but a

prerequisite for balanced deployment. It ensures that AI systems are documented in terms of execution time and memory usage, evaluated fairly across rare and common clinical categories, and capable of delivering precise, comprehensive classification across seven clinically relevant patient groups.

Biderstrand et al. [23] investigated a hybrid approach for classifying skin cancer that integrates convolutional neural networks (CNNs) with a closed-form continuous-time framework. The objective of their study was to address the shortcomings of techniques reliant solely on deep learning by merging the representational strengths of CNNs with the mathematical rigor of continuous-time modeling, thereby improving both accuracy and interpretability. Methodologically, CNNs were employed to extract features from dermoscopic images, while the continuous-time component added a structured analytical layer to enhance classification outcomes. This combined strategy aimed to synchronize data-driven learning with deterministic modeling, decreasing dependence on extensive datasets and alleviating risks associated with overfitting. The innovation lay in the hybridization process itself: by integrating statistical learning with closed-form solutions, the framework provided a new avenue for developing more resilient and theoretically sound diagnostic systems. Evaluation metrics included accuracy, sensitivity, specificity, and comparative assessments against conventional CNN baselines, with findings indicating improved classification stability and potential gains in generalization. From the standpoint of computational dermatology, several research gaps remain. At the dataset level, although dermoscopic images were utilized, the scale and diversity of the data were limited, restricting generalization across populations and skin tones. At the model level, while the hybrid approach introduced mathematical rigor, the absence of fully end-to-end deep learning architectures constrained adaptability, as the system relied on external formulations rather than integrated pipelines. At the metric level, conventional measures such as accuracy and specificity were reported, but fairness-aware and bias-sensitive metrics were absent, limiting evaluation across diverse patient groups. In this thesis, the introduction of Macro-F1 and Weighted-F1 scores advances evaluation beyond conventional binary measures, providing a more balanced and inclusive framework for multi-class classification across seven lesion categories. At the limitation level, the complexity of merging continuous-time formulations with deep learning frameworks required substantial computational resources and extensive validation, underscoring challenges in scalability and accessibility. In conclusion, Biderstrand et al. [23] underscored the significance of hybrid methodologies, demonstrating that the combination of CNN-based feature learning with mathematically grounded models can enhance diagnostic reliability, promote interpretability, and align with the overarching

objective of ensuring safe and clinically pertinent AI implementation in dermatology. Within the scope of this thesis, these findings reinforce the trajectory toward robust, scalable, and equitable diagnostic frameworks. By explicitly incorporating Macro and Weighted-F1 metrics for seven lesion categories, this research advances beyond prior work, offering a framework that improves diagnostic accuracy while ensuring fairness, generalizability, and accountability in AI-driven dermatology. In addition, M. M. Hossain et al. [24] introduced a maximum voting ensemble framework that integrates several state-of-the-art pre-trained deep learning models, including EfficientNet, DenseNet, and ResNet, for skin cancer detection. Their methodology emphasized uniform preprocessing, augmentation, and transfer learning strategies to ensure equitable use of backbones. The innovation lay in the ensemble design itself: majority voting provided resilience against biases of individual models, stabilized performance across lesion types, and reduced prediction variance. Evaluation metrics included accuracy, sensitivity, specificity, and AUC, with findings showing that the ensemble consistently outperformed individual models, thereby enhancing generalization and diagnostic consistency. From the perspective of computational dermatology, several research gaps remain. At the dataset level, reliance on dermoscopic collections limited diversity and external validation, restricting generalizability across populations and skin tones. At the model level, dependence on the diversity of base models introduced fragility, as ensemble strength was tied to the chosen architectures rather than a unified end-to-end framework. At the metric level, while conventional measures such as accuracy and AUC were reported, fairness-aware and bias-sensitive metrics were absent, and computational resource profiling (execution time, memory usage) was not explicitly addressed, limiting assessment of efficiency. In this thesis, the introduction of Macro-F1 and Weighted-F1 scores advances evaluation beyond conventional binary measures, providing a more balanced and inclusive framework for multi-class classification across seven lesion categories, while also documenting computational efficiency to ensure scalability. At the limitation level, the absence of multi-center datasets and resource profiling constrained scalability and clinical deployment, underscoring the necessity of standardized evaluation protocols that combine fairness with efficiency. In conclusion, Hossain et al. [24] demonstrated the promise of ensemble learning in dermatology, yet their work was bounded by dataset homogeneity, reliance on base model diversity, conventional evaluation metrics, and omission of computational profiling. Rather than serving merely as an aggregation experiment, their contribution highlights the necessity of embedding ensemble strategies as a design principle for clinically dependable AI. Within the scope of this thesis, the integration of maximum voting ensembles into broader systems is not only a methodological refinement but

a prerequisite for robust deployment. By explicitly incorporating Macro and Weighted-F1 metrics for seven lesion categories alongside execution time and memory usage profiling, this research advances beyond prior work, offering a framework that enhances fairness, generalizability, efficiency, and diagnostic reliability in AI-driven dermatology. Furthermore, D. Kim et al. [27] advanced a multi-view CNN framework for skin lesion classification, designed to leverage diverse perspectives of dermoscopic images. Their methodology developed CNN architectures capable of processing multiple views concurrently, thereby capturing richer contextual and structural details of lesions. The innovation lay in the integration of multi-view data, which enhanced feature representation and reduced the risk of overlooking critical diagnostic indicators. Evaluation metrics included accuracy, sensitivity, specificity, and AUC, with results showing that multi-view CNNs outperformed single-view models, particularly in handling complex lesion structures. Research gaps remain. At the dataset level, the absence of standardized multi-view dermoscopic datasets limited diversity and external validation. At the model level, the computational demands of multi-view CNNs constrained scalability and real-time deployment. At the metric level, while conventional measures were reported, interpretability-focused and efficiency-oriented metrics were absent, restricting systematic evaluation of clinical applicability. At the limitation level, the need for standardized multi-view datasets and the high resource requirements hindered reproducibility and broader adoption. The contribution of Kim et al. [27] demonstrates the potential of multi-view CNNs in dermatology, yet the study remains bounded by dataset scarcity, computational intensity, conventional evaluation metrics, and limited external validation. Rather than serving merely as an architectural extension, their work highlights the necessity of embedding multi-view perspectives as a design principle in clinical AI. For the purposes of this thesis, the integration of multi-view CNN frameworks into ensemble systems is not only a methodological refinement but a prerequisite for robust deployment. It ensures that AI systems are documented in terms of execution time and memory usage, balanced across rare and common clinical categories, and capable of delivering precise, comprehensive classification across seven clinically relevant patient groups. Moreover,

A. El-Sayed et al. [28] conducted a critical reassessment of the HAM10000 dataset, exposing issues of class imbalance, labeling inconsistencies, and the presence of artifacts. Their contribution focused on dataset innovation, proposing improved preprocessing and metadata documentation to enhance reliability. Evaluations centered on dataset characterization, distribution analysis, and baseline model performance, underscoring the risks of biased outcomes if HAM10000 is not carefully managed. Research gaps remain. At the dataset level,

restricted sources and lack of multi-center diversity limited generalizability. At the model level, reliance on baseline CNNs without advanced architectures constrained diagnostic robustness. At the metric level, conventional measures were reported, but fairness-aware and interpretability metrics were absent. At the limitation level, absence of external validation and artifact management strategies hindered reproducibility. In conclusion, the work of El-Sayed et al. [28] emphasizes the necessity of rigorous dataset evaluation as a foundation for dependable AI in dermatology. For this thesis, such reassessment ensures that ensemble systems are trained on balanced, artifact-controlled data, thereby enabling reproducible results and comprehensive classification across clinically relevant categories. Thus, L. Wang et al. [29] examined the application of Vision Transformers (ViTs) for skin lesion classification, marking a methodological shift from CNN-based frameworks to transformer-based designs. Their approach involved adapting ViTs to dermoscopic datasets with tailored preprocessing and augmentation, enabling the models to capture long-range dependencies and global contextual features. The innovation lay in demonstrating that transformers could match or surpass CNNs in medical imaging, offering robustness and interpretability. Evaluation metrics included accuracy, AUC, F1-score, and per-class performance, with results indicating competitive outcomes across lesion categories. Research gaps remain. At the dataset level, reliance on limited dermoscopic collections restricted diversity and external validation. At the model level, the computational demands of ViTs constrained scalability and real-time deployment. At the metric level, while conventional measures were reported, fairness-aware and efficiency-oriented metrics were absent, limiting holistic evaluation. At the limitation level, the absence of standardized multi-center datasets and high resource requirements hindered reproducibility and broader clinical adoption. The contribution of Wang et al. [29] illustrates the potential of transformer-based architectures in dermatology, yet the study remains bounded by dataset scarcity, computational intensity, conventional evaluation metrics, and limited external validation. Rather than serving merely as a technical exploration, their work highlights the necessity of embedding transformer models as a design principle in clinical AI. For the purposes of this thesis, the integration of Vision Transformers into ensemble systems is not only a methodological refinement but a prerequisite for robust deployment. It ensures that AI systems are documented in terms of execution time and memory usage, balanced across rare and common clinical categories, and capable of delivering precise, comprehensive classification across seven clinically relevant patient groups.

N. Patel et al. [30] presented a comparative study between CNN backbones and hybrid models for skin lesion classification, systematically analyzing trade-offs in accuracy,

sensitivity, and computational efficiency. Their methodology combined fine-tuned CNNs with hybrid designs that incorporated attention mechanisms and handcrafted features, offering nuanced insights into robustness and interpretability. The innovation lay in their transparent comparative reporting, which highlighted methodological diversity as a means of strengthening diagnostic pipelines.

Research gaps remain. At the dataset level, reliance on specific dermoscopic collections limited diversity and external validation. At the model level, hybrid designs introduced complexity and dependence on handcrafted components, constraining scalability. At the metric level, although accuracy, AUC, and macro-F1 were reported, fairness-aware and efficiency-oriented measures were absent. At the limitation level, lack of multi-center validation restricted reproducibility and clinical generalization. In conclusion, the work of Patel et al. [30] underscores the importance of methodological diversity, yet remains bounded by dataset scarcity, architectural complexity, conventional metrics, and limited external validation. For this thesis, the integration of hybrid approaches into ensemble systems is not only a methodological refinement but a prerequisite for reproducible deployment, ensuring balanced evaluation across rare and common lesion categories, documented efficiency in execution time and memory usage, and comprehensive classification across seven clinically relevant patient groups.

A. Bassel et al. [35] proposed a hybrid deep learning framework for automatic skin cancer classification, distinguishing malignant from benign lesions. Their approach integrated CNN-based feature extraction with hybrid layers, aiming to balance efficiency and diagnostic precision. The innovation lay in combining multiple learning paradigms to enhance reliability, with evaluation metrics including accuracy, sensitivity, specificity, and AUC. Results indicated superior performance compared to conventional CNNs. Research gaps remain. At the dataset level, limited diversity and absence of multi-center validation restricted generalizability. At the model level, hybrid integration introduced architectural complexity, raising concerns about scalability. At the metric level, reliance on conventional measures excluded fairness-aware and efficiency-oriented evaluations. At the limitation level, external validation was lacking, constraining reproducibility and clinical adoption. The contribution of Bassel et al. [35] demonstrates the promise of hybrid models, yet the study remains bounded by dataset scarcity, architectural complexity, conventional metrics, and limited external validation. Rather than serving merely as a technical enhancement, their work highlights the necessity of embedding hybrid frameworks as a design principle in clinical AI. For the purposes of this thesis, the integration of hybrid models into ensemble systems is not only a methodological refinement

but a prerequisite for robust deployment, ensuring balanced classification across rare and common categories, documented efficiency in execution time and memory usage, and comprehensive diagnostic reliability across seven clinically relevant patient groups.

Y. Wu et al. [36] provided a comprehensive review of deep learning approaches for skin cancer classification, synthesizing findings across diverse architectures, datasets, and evaluation techniques. Their analysis illuminated persistent challenges such as class imbalance, dataset bias, and insufficient external validation, while also pointing toward promising directions including ensemble learning and transformer-based models. The innovation lay in offering a panoramic overview that situated contemporary advancements while critically identifying gaps. Research gaps remain. At the dataset level, variability and bias across collections limited generalizability. At the model level, heterogeneity in architectures complicated reproducibility and hindered standardization. At the metric level, inconsistency in reporting practices restricted fair comparison and interpretability. At the limitation level, lack of external validation and uneven methodological rigor constrained clinical translation. In conclusion, the review of Wu et al. [36] underscores the necessity of methodological consistency and reproducibility in dermatology-focused AI. For this thesis, their work provides a foundational perspective, guiding the integration of ensemble and transformer-based strategies into clinically relevant pipelines, ensuring balanced evaluation across lesion categories, documented efficiency in execution time and memory usage, and alignment with diagnostic reliability across seven patient groups.

M. Reddirouthu et al. [39] conducted a performance assessment of CNN-based models for skin cancer detection and classification, benchmarking diverse architectures under standardized preprocessing and training protocols. Their innovation lay in transparent comparative evaluation, which revealed trade-offs between lightweight and complex CNNs. Results showed deeper CNNs achieving higher accuracy, while efficient models offered competitive outcomes with reduced computational costs. Research gaps remain. At the dataset level, reliance on specific dermoscopic collections limited diversity and external validation. At the model level, comparisons were restricted to CNNs, excluding hybrid or transformer-based alternatives. At the metric level, although accuracy and AUC were reported, fairness-aware and efficiency-oriented measures were absent. At the limitation level, lack of multi-center validation constrained reproducibility and clinical generalization. The contribution of Reddirouthu et al. [39] underscores the importance of balancing accuracy with efficiency, yet the study remains bounded by dataset scarcity, architectural exclusivity, conventional metrics, and limited external validation. Rather than serving merely as a benchmarking exercise, their

work highlights the necessity of embedding efficiency-aware CNN backbones into ensemble systems. For the purposes of this thesis, such integration is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and comprehensive diagnostic reliability across seven clinically relevant patient groups. In addition, V. Kumar and B. Sinha [42] investigated deep learning models based on transfer learning for dermoscopy image classification, employing pre-trained CNNs adapted to medical imaging tasks. Their methodology emphasized overcoming data scarcity by leveraging pre-trained weights, with results showing superior generalization compared to models trained from scratch. The innovation lay in demonstrating transfer learning as a viable strategy for medical imaging, with evaluation metrics including accuracy, sensitivity, and specificity confirming consistent improvements. Research gaps remain. At the dataset level, reliance on limited dermoscopic collections restricted diversity and external validation. At the model level, dependence on pre-trained weights sourced from non-medical domains constrained domain specificity. At the metric level, conventional measures were reported, but fairness-aware and efficiency-oriented metrics were absent. At the limitation level, lack of multi-center validation and computational profiling hindered reproducibility and scalability. The contribution of Kumar and Sinha [42] underscores the promise of transfer learning in dermatology, yet the study remains bounded by dataset scarcity, domain dependence, conventional metrics, and limited external validation. Rather than serving merely as a demonstration of transfer learning, their work highlights the necessity of embedding such strategies into ensemble systems. For the purposes of this thesis, transfer learning is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and comprehensive diagnostic reliability across seven clinically relevant patient groups. Furthermore, J. Li and M. Zhao [45] explored ResNet variants for melanoma detection, systematically fine-tuning architectures on dermoscopic datasets to enhance lesion classification. Their innovation lay in architectural optimization, showing that tailored ResNet variants can surpass standard implementations. Results demonstrated improvements in accuracy and robustness, positioning ResNet backbones as strong candidates for clinical pipelines. Research gaps remain. At the dataset level, reliance on limited dermoscopic collections restricted diversity and external validation. At the model level, optimization was confined to ResNet variants, excluding hybrid or transformer-based alternatives. At the metric level, conventional measures such as accuracy and sensitivity were reported, but fairness-aware

and efficiency-oriented metrics were absent. At the limitation level, lack of multi-center validation constrained reproducibility and clinical generalization. The contribution of Li and Zhao [45] underscores the potential of ResNet-based backbones in dermatology, yet the study remains bounded by dataset scarcity, architectural exclusivity, conventional metrics, and limited external validation. Rather than serving merely as an architectural refinement, their work highlights the necessity of embedding ResNet variants into ensemble systems. For the purposes of this thesis, such integration is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and comprehensive diagnostic reliability across seven clinically relevant patient groups. Moreover, V. Aruna et al. [47] introduced DeepSkin, a CNN-based framework for skin cancer classification, emphasizing preprocessing and augmentation to enhance generalization. Their innovation lay in creating a streamlined yet effective pipeline, evaluated through accuracy and AUC, with results showing competitive performance. Research gaps remain. At the dataset level, limited scale and lack of multi-center diversity restricted generalizability. At the model level, exclusive reliance on CNN architectures excluded hybrid or transformer-based alternatives. At the metric level, conventional measures dominated, omitting fairness-aware and efficiency-oriented evaluations. At the limitation level, absence of external validation constrained reproducibility and clinical translation. The contribution of Aruna et al. [47] highlights the promise of simplified CNN pipelines, yet remains bounded by dataset scarcity, architectural exclusivity, and conventional metrics. For the purposes of this thesis, integrating DeepSkin into ensemble systems is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and comprehensive diagnostic reliability across seven clinically relevant patient groups.

Thus, P. Natha and P. RajaRajeswari [50] proposed ensemble learning techniques for skin cancer prediction, combining multiple CNN architectures to enhance diagnostic dependability. Their innovation lay in demonstrating that ensemble models outperform individual networks in terms of robustness and stability. Evaluations included accuracy and class-wise performance, with findings showing improved equity across lesion categories. Research gaps remain. At the dataset level, reliance on limited dermoscopic collections restricted diversity and external validation. At the model level, ensemble strategies increased computational demands, raising concerns about scalability. At the metric level, conventional measures dominated, omitting fairness-aware and efficiency-oriented evaluations. At the

limitation level, high resource requirements and lack of multi-center validation constrained reproducibility and clinical adoption. The contribution of Natha and RajaRajeswari [50] underscores the promise of ensemble learning in dermatology, yet remains bounded by dataset scarcity, computational intensity, conventional metrics, and limited external validation. Rather than serving merely as a technical enhancement, their work highlights the necessity of embedding ensemble methods into clinically dependable AI pipelines. For the purposes of this thesis, ensemble integration is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and comprehensive diagnostic reliability across seven clinically relevant patient groups.

M. Hosseinzadeh et al. [53] developed a hybrid model for skin cancer detection, merging ensemble learning with deep learning to enhance accuracy and robustness. Their approach integrated multiple CNN architectures into an ensemble framework, leveraging complementary strengths while reducing variance. The innovation lay in the synergistic collaboration of ensemble and deep learning, which consistently outperformed individual CNNs. Evaluations included accuracy, sensitivity, specificity, and AUC, with results showing balanced performance across lesion classes.

Research gaps remain. At the dataset level, reliance on limited dermoscopic collections restricted diversity and external validation. At the model level, ensemble integration introduced computational demands that challenge scalability. At the metric level, conventional measures dominated, omitting fairness-aware and efficiency-oriented evaluations. At the limitation level, lack of multi-center validation constrained reproducibility and clinical adoption. The contribution of Hosseinzadeh et al. [53] underscores the promise of hybrid ensemble-deep learning methodologies, yet remains bounded by dataset scarcity, computational intensity, conventional metrics, and limited external validation. Rather than serving merely as a technical enhancement, their work highlights the necessity of embedding hybrid ensemble strategies into clinically dependable AI pipelines. For the purposes of this thesis, such integration is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and comprehensive diagnostic reliability across seven clinically relevant patient groups.

D. A. Okuboyejo and O. O. Olugbara [56] introduced a weighted majority voting ensemble for skin lesion classification, aggregating predictions from multiple deep learning models. Their innovation lay in the weighted voting mechanism, which enhanced reliability by

assigning weights based on model performance. Results demonstrated improved accuracy and more balanced classification outcomes, underscoring the potential of ensemble strategies in dermatology. Research gaps remain. At the dataset level, dependence on specific dermoscopic collections limited diversity and external validation. At the model level, reliance on weighted tuning introduced dataset-specific sensitivity, constraining scalability. At the metric level, conventional measures such as accuracy dominated, omitting fairness-aware and efficiency-oriented evaluations. At the limitation level, absence of multi-center validation restricted reproducibility and clinical adoption. The contribution of Okuboyejo and Olugbara [56] highlights the efficacy of weighted ensembles in mitigating biases inherent to individual models, yet remains bounded by dataset scarcity, tuning dependence, conventional metrics, and limited external validation. For the purposes of this thesis, embedding weighted ensemble mechanisms into ensemble frameworks is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and comprehensive diagnostic reliability across seven clinically relevant patient groups.

A. R. Maurya et al. [58] proposed a dual autoencoder framework enhanced with attention mechanisms and integrated extreme learning machines for skin cancer detection. Their innovation lay in coupling deep latent feature extraction with efficient classifiers, yielding competitive accuracy and resilience. Attention mechanisms guided feature representation, while extreme learning machines accelerated classification efficiency. Research gaps remain. At the dataset level, reliance on specific dermoscopic collections restricted diversity and external validation. At the model level, architectural sophistication introduced complexity that may hinder scalability. At the metric level, conventional measures such as accuracy and sensitivity dominated, omitting fairness-aware and efficiency-oriented evaluations. At the limitation level, dependence on dataset-specific tuning constrained reproducibility and clinical adoption. The contribution of Maurya et al. [58] underscores the importance of attention-driven feature learning, yet remains bounded by dataset scarcity, architectural complexity, conventional metrics, and limited external validation. For the purposes of this thesis, embedding attention-guided autoencoders within ensemble frameworks is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and comprehensive diagnostic reliability across seven clinically relevant patient groups.

K. Sethanan et al. [59] presented a dual AMIS-ensemble deep learning framework for skin cancer classification, integrating multiple deep learning models within an ensemble optimized by the Artificial Multiple Intelligence System (AMIS). Their innovation lay in the optimization mechanism, which enhanced diagnostic precision and reliability, yielding performance superior to individual models. Research gaps remain. At the dataset level, reliance on limited dermoscopic collections restricted diversity and external validation. At the model level, ensemble optimization introduced computational intensity that may hinder scalability. At the metric level, conventional measures such as accuracy and AUC dominated, omitting fairness-aware and efficiency-oriented evaluations. At the limitation level, absence of multi-center validation constrained reproducibility and clinical adoption. The contribution of Sethanan et al. [59] underscores the promise of sophisticated ensemble optimization techniques, yet remains bounded by dataset scarcity, computational demands, conventional metrics, and limited external validation. For the purposes of this thesis, embedding AMIS-optimized ensembles into clinical AI pipelines is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and contribute to broader diagnostic reliability across seven clinically relevant patient groups. In addition, M. H. Imam et al. [63] advanced skin cancer classification by proposing a deep learning ensemble that fused DenseNet and MobileNet architectures. Their methodological contribution lay in leveraging DenseNet’s hierarchical feature propagation with MobileNet’s lightweight efficiency, thereby achieving a synergistic balance between representational depth and computational feasibility. Reported outcomes demonstrated improved diagnostic accuracy and robustness compared to single-model baselines, with evaluations encompassing accuracy and lesion-specific performance metrics. Nonetheless, several research gaps persisted. At the dataset level, reliance on dermoscopic archives limited adaptability to heterogeneous clinical environments. At the model level, the fusion approach, while effective, introduced architectural complexity that challenged scalability and interpretability. At the metric level, conventional accuracy-driven assessments dominated, overlooking fairness-aware and efficiency-oriented measures. At the limitation level, absence of multi-center validation constrained reproducibility and hindered clinical translation. The contribution of Imam et al. [63] highlights the promise of CNN-based ensembles in dermatology, yet remains bounded by dataset dependence, architectural complexity, conventional metrics, and limited external validation. In future work, extending such fusion-based ensembles into broader frameworks that incorporate multiple pre-trained backbones and efficiency profiling may provide more balanced classification across

rare and common lesion categories, enhance diagnostic reliability, and support clinically viable deployment across diverse patient groups. Furthermore, M. A. Arshed et al. [64] explored multi-class skin cancer categorization by combining vision transformers (ViTs) with CNN-based pre-trained models. Their innovation lay in the hybridization of CNNs for local feature extraction and ViTs for capturing global contextual relationships, yielding superior performance compared to standalone models. Assessments included accuracy, AUC, and class-specific metrics, with results confirming enhanced generalization and diagnostic accuracy. Research gaps remain. At the dataset level, reliance on large dermoscopic datasets restricted applicability in data-scarce environments. At the model level, hybrid CNN–ViT architectures introduced considerable computational demands, challenging scalability. At the metric level, conventional measures dominated, omitting fairness-aware and efficiency-oriented evaluations. At the limitation level, absence of multi-center validation constrained reproducibility and clinical adoption. The contribution of Arshed et al. [64] underscores the potential of CNN–transformer hybrids in dermatology, yet remains bounded by dataset dependence, computational intensity, conventional metrics, and limited external validation. For the purposes of this thesis, embedding CNN–ViT hybrids into ensemble frameworks is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and contribute to broader diagnostic reliability across seven clinically relevant patient groups. Moreover, Gautam et al. [65] introduced FusionEXNet, an interpretable deep learning framework for skin cancer detection that integrates multiple feature extraction pathways into a unified model. Their innovation lay in balancing predictive accuracy with transparency, enabling clinicians to better understand model decisions. Evaluations included accuracy, sensitivity, specificity, and interpretability assessments, with results showing competitive performance and improved fidelity in explanations. Research gaps remain. At the dataset level, reliance on limited dermoscopic collections restricted diversity and external validation. At the model level, architectural complexity introduced challenges for scalability and deployment. At the metric level, conventional measures dominated, omitting fairness-aware and efficiency-oriented evaluations. At the limitation level, absence of multi-center validation constrained reproducibility and clinical adoption. The contribution of Gautam et al. [65] underscores the importance of interpretable deep learning frameworks in dermatology, yet remains bounded by dataset scarcity, architectural complexity, conventional metrics, and limited external validation. For the purposes of this thesis, embedding interpretable models such as FusionEXNet into ensemble pipelines is not only a

methodological refinement but a prerequisite for trustworthy AI deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and contribute to broader diagnostic reliability across seven clinically relevant patient groups. Thus, I. A. Kandhro et al. [66] conducted a performance evaluation of the E-VGG19 model, an optimized variant of VGG19 tailored for real-time skin cancer detection and classification. Their innovation lay in architectural refinements that balanced computational efficiency with diagnostic precision, making the model suitable for real-time clinical deployment. Evaluations included accuracy, sensitivity, specificity, and latency, with results confirming improved applicability in practice. Research gaps remain. At the dataset level, reliance on specific dermoscopic collections restricted diversity and external validation. At the model level, optimization was confined to CNN variants, excluding hybrid or transformer-based alternatives. At the metric level, conventional measures dominated, omitting fairness-aware and efficiency-oriented evaluations. At the limitation level, absence of multi-center validation constrained reproducibility and broader clinical adoption. The contribution of Kandhro et al. [66] underscores the importance of efficient CNN variants for real-world dermatology, yet remains bounded by dataset scarcity, architectural exclusivity, conventional metrics, and limited external validation. For the purposes of this thesis, embedding optimized CNNs such as E-VGG19 into ensemble frameworks is not only a methodological refinement but a prerequisite for robust deployment, ensuring balanced classification across rare and common lesion categories, documented efficiency in execution time and memory usage, and contribute to broader diagnostic reliability across seven clinically relevant patient groups.

Table (2.1): Summary of Skin Cancer Detection Approaches: Traditional, CNN, Ensemble and Transformer Models.

Approach	Study	Methodology	Contribution	Limitations
Traditional	<u>Fitzpatrick et al. [1]</u>	Classification of skin into six sun-reactive types (I–VI)	Established a global standard for assessing skin cancer risk and response to laser treatment	Does not directly analyze lesion images; limited generalizability across ethnic groups
Traditional	<u>Kittler et al.[2]</u>	Evaluation of diagnostic accuracy using dermoscopy vs. naked-eye examination	Demonstrated that dermoscopy significantly improves diagnostic accuracy and reduces errors	Small dataset; did not leverage large-scale archives like HAM10000
CNN	Sethanan et al. [20]	MobileNet-Lite framework tested on small medical datasets using standard classification metrics	Demonstrated energy-efficient and sustainable deep learning for medical image classification	Limited datasets, no ensemble/advanced models, narrow metrics (accuracy/F1 only), no clinical validation
CNN	V. Aruna et al. [47] (DeepSkin)	Simplified CNN framework with preprocessing enhancements	Simple design with competitive results	Limited dataset and absence of external validation
CNN	I. A. Kandhro et al. [66] (E VGG19)	Optimized VGG19 architecture for real-time detection and classification	Balanced computational efficiency with diagnostic accuracy	Dataset dependence and lack of multi-center validation
Ensemble	Hossain et al.[24] (Max Voting Ensemble)	Integration of EfficientNet, DenseNet, ResNet via majority voting	Improved generalization, stabilized performance across lesion types, reduced prediction variance	Dependent on diversity of base models; lacked computational profiling (execution time, memory usage)
Ensemble	R. Natha et al. [50]	Multiple CNNs combined via ensemble techniques	Improved robustness and stability	High computational requirements and limited external validation
Ensemble	M. Hosseinzadeh et al. [53]	Hybrid ensemble integrating CNN architectures	Balanced performance across lesion classes and reduced variance	Computationally intensive and restricted generalization

Approach	Study	Methodology	Contribution	Limitations
Ensemble	D. A. Okuboyejo and O. O. Olugbara [56]	Weighted majority voting ensemble	Reduced bias and enhanced reliability	Sensitive to dataset-specific tuning and absence of external validation
Ensemble	K. Sethanan et al. [59] (AMIS-ensemble)	Ensemble optimized with Artificial Multiple Intelligence System (AMIS)	Enhanced precision and reliability through optimization	Computationally demanding and lack of multi-center validation
Ensemble	B. Hassan et al. [63]	Multi-class ensemble of CNN classifiers	Increased diagnostic confidence and reduced bias	High computational cost and limited external validation
Transformers	Zhang et al., 2025 (Hybrid CNN–Transformer)	Combined CNN feature extraction with Transformer-based contextual modeling	Enhanced sensitivity for malignant classes; improved overall accuracy	Weakness in rare lesion categories; absence of multi-center external validation
Transformers	M. A. Arshed et al. [64]	Hybrid CNN–Vision Transformer framework	Synergistic feature representation combining local and global contexts	Requires large datasets and is computationally heavy
Transformers / Interpretable AI	Y. Gautam et al. [65] (FusionEXNet)	Interpretable deep learning with fused feature extraction pathways	Balanced predictive accuracy with transparency for clinicians	Architectural complexity and need for broader validation
Hybrid Attention	A. R. Maurya et al. [58]	Dual autoencoder with attention mechanisms and Extreme Learning Machines	Improved feature representation through attention guidance	Complex architecture and dataset-specific reliance

CHAPTER 3:

METHODOLOGY

3.1 Introduction

This chapter provides The proposed methodology employs a Multi-DCNN ensemble with Max-Voting to advance dermoscopic skin lesion classification, ensuring robustness, reproducibility, and fairness in evaluation. It begins with a dataset overview (HAM10000) and its provenance and selection rationale, proceeds through preprocessing and stratified splitting, details the chosen pre-trained architectures and their properties and motivations, explains the Max-Voting fusion and its justification, and finally describes imbalance handling, training controls, and evaluation metrics.

3.2 Methodology

This architecture utilizes five pre-trained models (DenseNet121, EfficientNet-B0, ResNet50, VGG19, MobileNetV2), which are carefully balanced for their representational capabilities and computational efficiency. The outputs are integrated through a Max-Voting strategy to generate trustworthy predictions. In order to tackle class imbalance, three strategies are applied:

WeightedRandomSampler, FocalLoss, and Alpha_tensor, all focused on enhancing sensitivity to less common categories like melanoma. This thesis provides a thorough methodology for the classification of dermoscopic skin lesions using deep learning, leveraging the HAM10000 dataset, which contains a wide array of skin images across seven diagnostic categories. The approach initiates with preprocessing, involving image resizing, normalization, and the partitioning of data into training, validation, and testing subsets, utilizing mild augmentations to bolster model generalization. In summary, this methodology integrates rigorous data preparation, architectural diversity, ensemble fusion via Max-Voting, advanced imbalance handling, and carefully tuned training settings. Together, these components ensure robust learning, reduced overfitting, and improved validation accuracy through systematic reduction of validation loss.

3.3 Methodological Framework of the Proposed Approach

The primary objective of this research is to investigate whether the proposed Max-Voting-based ensemble model can more accurately seven-class prediction than the individual models. The essential stages of the proposed methodology are outlined below as depicted in Figure (3.1):

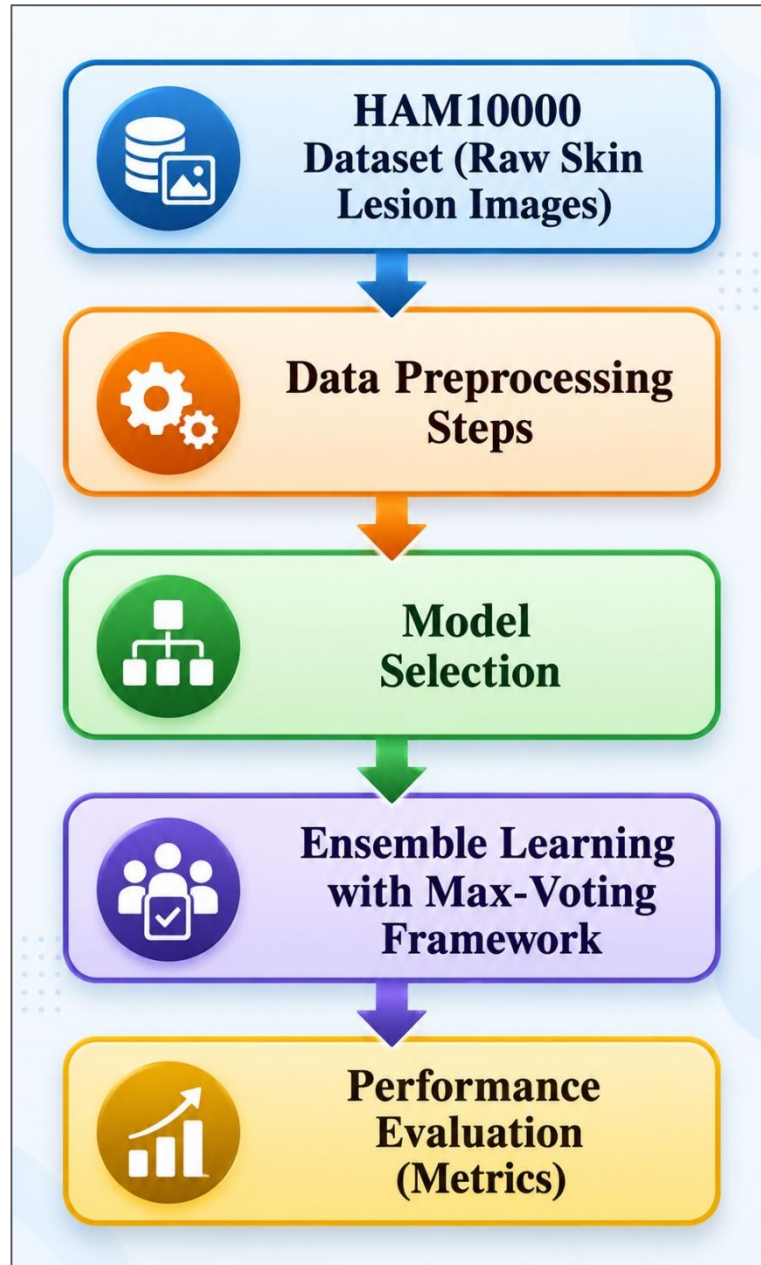


Figure (3.1): Essential Stages of the Proposed Methodology.

The essential stages of the Main Details of Proposed Methodology are outlined below:

1. Dataset Overview

HAM10000 Dataset: Utilize the HAM10000 dataset "Human Against Machine with 10000 training images" dataset, a comprehensive collection of dermoscopic images with annotations for various skin lesion types. HAM10000 dataset includes seven classes of skin lesions are as follows:(Melanoma, Melanocytic nevus, Basal cell carcinoma, Actinic keratosis, Benign keratosis, Dermatofibroma, Vascular lesion). This dataset contains 10,015 images of skin lesions, classified into two categories: benign (7,481images) and malignant (2,534

images), the Figure (3.2) visually presents sample images depicting both malignant and benign cases from the HAM10000 dataset.

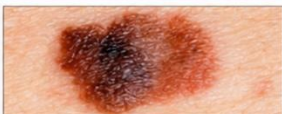
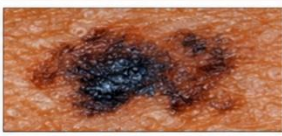
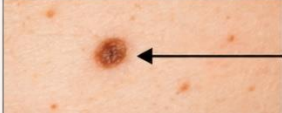
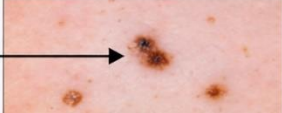
	Benign	Malignant	
Symmetrical			Asymmetrical (the two sides do not match)
Borders are even			Borders are uneven
One color			Two or more colors
Smaller than 1/4 inch			Larger than 1/4 inch
Ordinary mole			Changing in size, shape, color, or another trait

Figure (3.2): Skin Cancer Malignant and benign sample images. [69].

The HAM10000 dataset is an excellent choice for skin cancer detection research using Multi-DCNN models due to its large size, diversity, high-quality images, seven classes of skin lesions, class imbalance, annotations, dermatoscopic images, and availability for benchmarking and comparison.

2. Data Preprocessing Steps

– Image Resizing

Purpose: Standardize the input dimensions for all images.

Technique: Resize all images to a uniform size, such as 224x224 pixels, which is compatible with most DCNN architectures.

– **Data Augmentation:** Apply techniques such as rotation, flipping, and zooming, and to increase dataset variability and prevent overfitting.

– **Normalization:** Purpose scale pixel values to a consistent range to improve model convergence.

Technique: Normalize pixel values to a range of [0, 1] or standardize to have a mean of 0 and a standard deviation of 1.

– ***Splitting the Dataset***

Stratified Split: Stratified splitting was applied at the image level to maintain balanced representation of all lesion categories, including rare classes.

Step 1: The dataset was initially divided into two main subsets, with 80% allocated for training and 20% reserved for testing.

Step 2: The training subset was further split into 80% final training and 20% validation. Final proportions Training: 64%, Validation: 16%, Testing: 20%.

3. Model Selection

Combine heavy and lightweight architectures to balance representational richness and computational efficiency, leveraging complementary feature spaces. And utilizing an ensemble approach allows for the strengths of each model to compensate for the weaknesses of others, leading to more reliable predictions. Utilize the following pre-trained models for feature extraction: (DenseNet121, EfficientNetB0, ResNet50, VGG19, MobileNetV2).

4. Ensemble Learning with Max-Voting Approach

after obtaining predictions from each model, implement a max-voting mechanism for each test sample and select the class that receives the highest number of votes from the individual models.

5. Performance Evaluation (Metrics): Evaluate the performance using metrics such as precision, recall, F1-score, Overall Accuracy, Macro-F1 and Weighted-F1 scores.

3.4 Detailed View of Proposed Algorithm

The process of detection of seven-class prediction starts with image data collection, followed by steps such as image preprocessing, data augmentation, feature extraction through, the five proposed individual pre-trained models (DenseNet121, EfficientNet-B0, ResNet50, VGG19, MobileNetV2), and finally, classification by combining the outputs of individual pre-trained models through the proposed max voting-based ensemble model. Transfer learning was used in our models to reduce the computation time and encourage better results and performance.

This section outlines the detailed methodology we used to train and test our CNN models and predict the final outcome using the proposed Max-Voting-based model as shown in Figure (3.3).

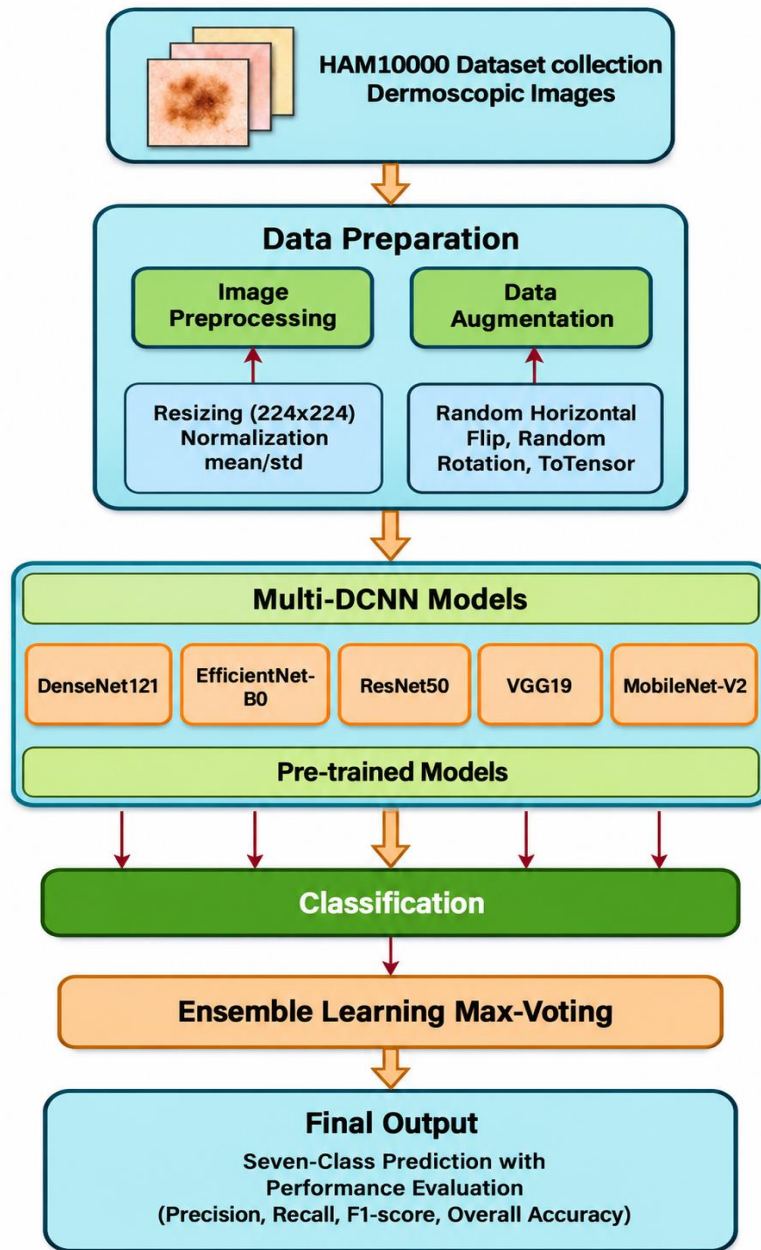


Figure (3.3): Block Diagram of Detailed View of Proposed Algorithm.

Figure (3.3) represents the entire workflow of the proposed Algorithm designed for a skin lesion detection system. This workflow incorporates a max-voting ensemble strategy for the classification of skin cancer utilizing pre-trained deep learning models. The process initiates with the selection of a variety of models, proceeds with the preprocessing of input images, the loading of the selected models, the generation of predictions for each model, the aggregation of predictions through majority voting, the evaluation of ensemble accuracy, and concludes with the display of the results. Algorithm depicts the data flow diagram for the proposed method, as illustrated in Figure (3.3).

Algorithm: Multi-DCNN Ensemble Classification Using Max -Voting for Skin Lesion Diagnosis

Input

- InputImages: Set of dermoscopic skin lesion images from HAM10000.
- SelectedModels: List of pre-trained CNN architectures [DenseNet121, EfficientNet-B0, ResNet50, VGG19, MobileNetV2].
- GroundTruthLabels: True class labels for evaluation.

Output

- EnsemblePredictions: Final predicted class for each image (seven-class prediction).
- PerformanceMetrics: accuracy, precision, recall, and F1-score.

Procedure

Step 1: Preprocessing and Augmentation Operations for Dermoscopic Images.

- Normalize (Training & Validation)
- Resize image to match model input dimensions (e.g., 224×224).
- Apply geometric transforms: Random Horizontal Flip, Random Rotation, ToTensor.

Step 3: Splitting the Dataset

- Split into 80% final training and 20% validation.
- Final proportions Training: 64%, Validation: 16%, Testing:20%.

Step 4: Class Imbalance Handling

- Apply WeightedRandomSampler sampling to maintain class distribution.
- Use FocalLoss and Alpha_tensor during training.

Step 5: Load Pre-Trained Models

- LoadModels←LoadModels(SelectedModels).
- Use optimizer like AdamW with learning rate scheduling.

Step 6: Generate Predictions

- For each image, predict class using all models.
- Store predictions from each model.

Step7: Max-Voting Ensemble

- Apply majority voting across model predictions.
- Select the most frequent class as final prediction.

Step 8: Final Classification

- Map output to seven-class prediction.

Step 9: Evaluate Performance

- Calculate precision, recall, F1-score, Overall Accuracy, Macro-F1 and Weighted-F1 scores.
- Generate Confusion Matrix.

Step 9: Display Results.

- The Essential Stages of the View Detailed of Proposed Algorithm are outlined below:
-

3.4.1 The HAM10000 Dataset

The HAM10000 dataset "Human Against Machine with 10000 training images" dataset, comprehensive collection of dermoscopic images with annotations for various skin lesion types. a dermoscopic collection derived from the ISIC-2018 challenge of the International Skin Imaging Collaboration, which aims to standardize and openly share skin imaging data for melanoma research. HAM10000 dataset includes seven classes of skin lesions are as follows: melanocytic nevi (nv) or nevus (nv), melanoma (mel), benign keratosis-like lesions (bkl), basal cell carcinoma (bcc), actinic keratoses carcinoma (akiec), vascular lesions (vasc), and dermatofibroma (df), are depicted in Figure (3.4).

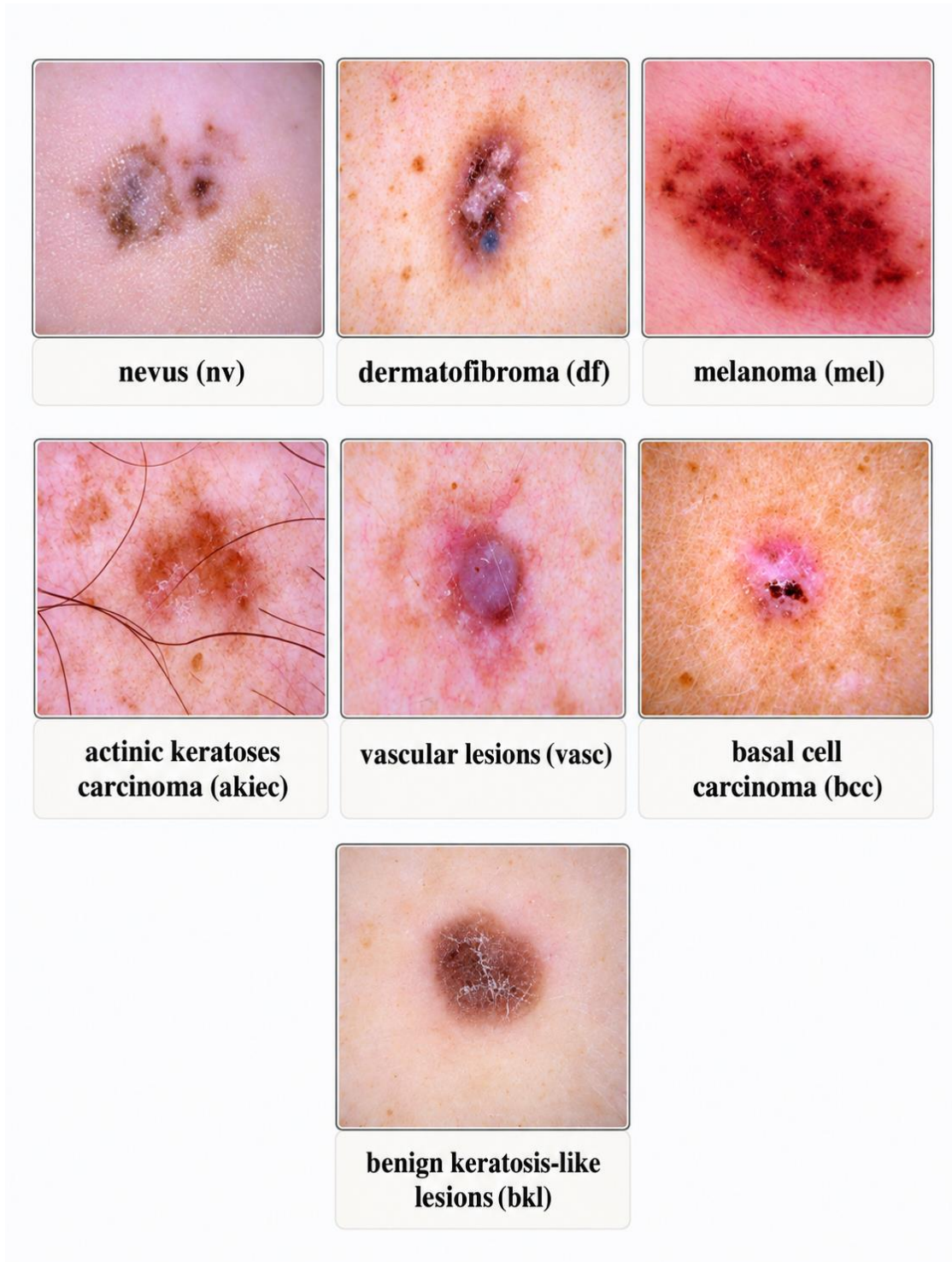


Figure (3.4): Seven Classes of Skin Lesions of HAM10000 Dataset. [70].

The HAM10000 dataset comprises seven different classes as described above and the class-wise categorization for the number of images is stated in Table (3.1). The distribution of images seems to be imbalanced. There are several distinctive advantages that render

HAM10000 the most appropriate choice for this study. These advantages can be summarized in the following points:

- ***Widely recognized benchmark:*** HAM10000 is among the most frequently employed datasets in dermoscopic image classification, thereby ensuring comparability with contemporary international research.
- ***Appropriate size and diversity:*** The dataset contains 10,015 dermoscopic images organized into seven diagnostic categories, providing sufficient data for training deep learning models while also reflecting the diversity encountered in clinical environments.
- ***Clinical realism:*** The images are collected from various clinical centers and devices, introducing variability that enhances the models' applicability to real-world scenarios.
- ***Class imbalance representation:*** HAM10000 comprises both rare yet clinically important classes (like melanoma, akiec, and dermatofibroma) and majority classes (like nevus).
- ***Connection to ISIC-2018 challenge:*** As part of the International Skin Imaging Collaboration, HAM10000 adheres to global standards and facilitates direct comparison with state-of-the-art research outcomes.
- ***Limitations:*** Despite the strengths of the proposed method, several limitations should be acknowledged. First, the study relied exclusively on the HAM10000 dataset, without incorporating external validation, which restricts generalizability. Second, no Grad-CAM visualization was applied, limiting interpretability of the model's decision process. Third, the absence of statistical significance testing reduces the robustness of comparative claims. Additionally, the dataset suffers from class imbalance and pronounced visual similarity between nevus and melanoma, which complicates classification. Finally, the melanoma F1-score remained relatively low, reflecting the challenge of distinguishing malignant lesions from visually similar benign cases. These limitations highlight areas for improvement in future work, particularly through external validation, explainability techniques, and statistical evaluation.

Table (3.1): The HAM10000 Dataset – Sample Distribution of Classes.

Class	Abbreviation	Number of Samples	Percentage of Dataset	Academic Note
Melanocytic Nevi	NV	6705	~67%	Dominates the dataset; provides strong representation of benign lesions but causes imbalance.
Melanoma	MEL	1113	~11%	Critical malignant class; relatively underrepresented compared to NV, requiring imbalance handling.
Benign Keratosis-like Lesions	BKL	1099	~11%	Common benign lesions; similar frequency to melanoma, useful for comparative classification.
Basal Cell Carcinoma	BCC	514	~5%	Malignant but less frequent; important for clinical detection.
Actinic Keratoses	AKIEC	327	~3%	Precancerous lesions; minority class requiring weighting strategies.
Vascular Lesions	VASC	142	~1%	Rare class; contributes to severe imbalance, often overlooked without sampling techniques.
Dermatofibroma	DF	115	~1%	Very rare benign lesion; critical for demonstrating model sensitivity to minority classes.

3.4.2 Dataset Preprocessing and Splitting

– *Image Standardization and Transformations*

In order to ensure compatibility with pre-trained deep convolutional neural networks (DCNNs), all images were resized to a consistent dimension (e.g., 224×224 pixels) and

standardized through Resize/Center-Crop operations. Normalization was conducted using the ImageNet mean/std and scale pixel values to a consistent range to improve model convergence. Which is essential for alignment with commonly used pre-trained architectures. As depicted in Table (3.2). In addition, the frequent application of geometric transformations such as rotation and flipping, can result in significant distortion of dermoscopic images of skin lesions. Such distortion contributes to the loss of subtle details and introduces considerable variability during the training process, culminating in erroneous training results.

• **Data Augmentation**

Beyond resizing and normalization, light and clinically appropriate augmentations were implemented. These augmentations included Resizing, Horizontal Flip Rotation, ToTensor, Normalize. As depicted in Table (3.2). Such enhancements increased the variability of the dataset and enhanced generalization while maintaining lesion morphology and preventing distortion of clinically significant features.

Table (3.2): Preprocessing and Augmentation Operations for Dermoscopic Images.

Processing Type	Dataset Phase	Mode / Range
Resizing	Training & Validation	224 × 224
Random Horizontal Flip	Training only	Probability = 0.5
Random Rotation	Training only	±10 degrees
ToTensor	Training & Validation	Converts image to PyTorch tensor format
Normalize	Training & Validation	Mean = [0.485, 0.456, 0.406] Std = [0.229, 0.224, 0.225]

– **Splitting the Dataset**

To ensure compatibility with neural network models, categorical diagnostic labels (dx) were first converted into numeric form using a LabelEncoder. This transformation facilitated efficient model training and evaluation. A two-stage stratified splitting strategy was employed to preserve per-class proportions across all subsets. In the first stage, the dataset was divided into training (80%) and testing (20%) subsets, with stratification on label and a fixed random_state = 42 to guarantee reproducibility. In the second stage, the training subset was

further partitioned into final training (80%) and validation (20%) sets, again under stratification. This procedure yielded final proportions of 64% training, 16% validation, and 20% testing. Importantly, this approach maintained near-balanced representation of rare but clinically significant classes, such as melanoma (mel), actinic keratosis (akiec), and dermatofibroma (df). Stratified splitting was applied at the image level to maintain balanced representation of all lesion categories, including rare classes. This approach treats each image as an independent diagnostic unit, ensuring proportional class distribution across training, validation, and test sets, thereby supporting reliable and unbiased model evaluation.

Validation and testing subsets: Only resizing and normalization were applied, thereby ensuring unbiased evaluation without augmentation-induced variability.

Data source link in code: In this study, the HAM10000 dataset was accessed directly via Kaggle (skin-cancer-mnist-ham10000) (kaggle.com in Bing), providing dermoscopic images alongside a CSV file containing essential metadata such as diagnosis, patient age, sex, and lesion localization. The integration of clinical variables with image data enables comprehensive medical analysis. Furthermore, Kaggle’s complimentary GPU and memory resources facilitate efficient training of complex deep learning models without reliance on costly local infrastructure.

3.4.3 Selection of pre-trained models

All five pre-trained models employed in the proposed method were initialized with ImageNet-pre-trained weights and configured with an input size of $224 \times 224 \times 3$. The final classification layers were modified to output seven classes, incorporating dropout and fully connected layers to enhance generalization. Each model operates with a suitable number of parameters in this proposed approach, ensuring a balance between computational efficiency and diagnostic accuracy. The training strategy adopted full fine-tuning without freezing early layers, allowing all parameters to be updated during optimization. As depicted in Table (3.3).

– DenseNet121(Dense connectivity Network 121-layer Network)

Advantages: DenseNet121 leverages a dense connectivity framework where each layer is connected to all subsequent layers, promoting feature reuse and smoother gradient flow. This design reduces the number of parameters compared to similarly deep networks while maintaining strong performance in medical imaging tasks. Hossain et al. [24] demonstrated

that DenseNet121, when integrated into an ensemble of pre-trained models, contributed significantly to improved diagnostic accuracy in skin cancer detection.

Architecture: The architecture resembles conventional CNNs, with alternating convolutional and pooling layers, followed by a global average pooling layer and a fully connected layer for classification. What distinguishes DenseNet121 is its dense connectivity mechanism, which enables efficient learning of intricate features while minimizing redundancy.

Limitations: The concatenation operations inherent in dense connectivity introduce memory overhead, and the model can be sensitive to batch size selection during training.

Prior Use: DenseNet121 has been widely applied in radiology and dermoscopy classification, consistently demonstrating reliable performance in medical image analysis.

Why Chosen: DenseNet121 offers an excellent balance between depth and performance, providing nuanced feature extraction with fewer parameters. Its proven success in medical imaging and its contribution to ensemble frameworks, as highlighted in Hossain et al. [24], make it a strong candidate for robust skin cancer classification.

– **EfficientNet-B0(Efficient Network B0 version)**

Advantages: EfficientNet-B0 employs a compound scaling strategy that balances depth, width, and resolution, yielding high accuracy per FLOP and making it ideal for resource-constrained environments. This efficiency allows the model to achieve competitive performance in medical imaging tasks while maintaining relatively low computational cost. Mehta et al. [25] demonstrated that transfer learning with EfficientNet-B0 provided strong results in skin lesion diagnosis, highlighting its suitability for clinical applications.

Architecture: The baseline EfficientNet-B0 integrates inverted residual blocks with squeeze-and-excitation optimization, followed by global average pooling and a fully connected classification layer. Its systematic scaling framework distinguishes it from conventional CNNs, enabling nuanced feature extraction with optimized efficiency.

Limitations: As the smallest variant in the EfficientNet family, B0 may trail larger versions (e.g., B4–B7) in peak accuracy. It can also be sensitive to hyperparameter tuning, particularly input resolution and batch size.

Prior Use: EfficientNet-B0 has been widely adopted in medical image classification, including dermatology and radiology. In Mehta et al [25], it was successfully applied to skin lesion datasets, achieving high diagnostic accuracy with reduced computational demand.

Why Chosen: EfficientNet-B0 offers a strong baseline with fast training and low compute requirements, making it an excellent choice for ensemble frameworks and real-world medical imaging scenarios where efficiency and reliability are critical.

– ResNet50 (Residual Connections Network 50-layers)

Advantages: ResNet50 introduces residual connections, which allow gradients to bypass certain layers and flow directly through the network. This design mitigates the vanishing gradient problem, enabling the training of deeper architectures with improved stability. As a result, ResNet50 is considered a robust and well-studied baseline in computer vision. Hossain et al. [24] highlighted its effectiveness when integrated into an ensemble of pre-trained models for skin cancer detection, where residual connections contributed to enhanced diagnostic accuracy.

Architecture: The structure of ResNet50 resembles conventional CNNs with convolutional and pooling layers, followed by fully connected layers for classification. Its distinctive feature is the skip (residual) connections, which add the input of a layer directly to its output, facilitating efficient learning and deeper representation.

Limitations: ResNet50 requires moderate to heavy computational resources and careful tuning of hyperparameters such as the learning rate. Without proper optimization, training may become unstable or inefficient.

Prior Use: ResNet50 has been widely adopted across image classification tasks, including medical imaging domains such as radiology and dermoscopy. Its proven reliability makes it a ubiquitous standard in deep learning applications.

Why Chosen: ResNet50 provides a reliable foundation and adds architectural diversity to ensemble frameworks. Its residual connections enhance training stability and feature extraction, making it a valuable component in multi-model strategies like the Max-voting ensemble employed by Hossain et al. [24].

– **VGG19 (Visual Geometry Group 19-layer Network)**

Advantages: Simple, consistent 3×3 stacks; a canonical benchmark.

Limitations: Heavy head and many parameters, less modern than EfficientNet-B0 and MobileNetV2. As depicted in Table (3.3).

Lacks Batch_Normlization in original blocks, increasing gradient variance and (lr) sensitivity.

Often starts with lower validation accuracy (~12–16%) than modern nets.

Prior use: Historical baseline in computer vision and academic comparisons.

Why chosen: The proposed methodology highlights training challenges on rare classes and the benefits of schedulers and regularization. Although VGG19 is sensitive to learning rate variations and requires extended training epochs, its inclusion is intentional, as its deeper and heavier architecture provides a stricter environment to mitigate the overconfidence induced by the Focal Loss function. This deliberate choice ensures that the ensemble framework evaluates both lightweight and heavyweight models, thereby enhancing robustness and reproducibility.

– **MobileNetV2 (Mobile Network Version 2 ~53 layers):**

Advantages: MobileNetV2 is distinguished as a neural network architecture specifically designed for efficient and rapid image classification on mobile and resource-constrained devices. By employing depthwise separable convolutions and linear bottlenecks, it reduces computational complexity while maintaining competitive performance. This balance between model size and accuracy makes MobileNetV2 highly suitable for real-time diagnosis and classification tasks, including dermoscopic skin cancer detection, where efficiency and reliability are critical [64].

Limitations: Lower representational capacity than heavy nets; may need augmentation/ensembles for fine distinctions.

Prior use: Frequent in resource-constrained medical applications.

Why chosen: Sethanan et al. [20] demonstrated the potential of lightweight CNNs for medical imaging through MobileNet-Lite-Health, yet their framework faced limitations in dataset diversity, metric scope, and clinical validation. Building on these insights, this thesis adopts MobileNetV2 as a more suitable option, offering enhanced representational capacity while preserving efficiency, which makes it an appropriate choice for the proposed method aiming to balance accuracy, scalability, and practical applicability in medical image analysis.

Table (3.3): CNN Model Configurations and Parameters.

<u>Model</u>	<u>Number of Parameters</u>	<u>Batch-size</u>	<u>Patience</u>	<u>Factor</u>	<u>Random seed</u>	<u>Num_workers</u>	<u>Criterion</u>	<u>Functions/Notes</u>
DenseNet121	8.1M	30 epochs	3	0.5	42	2	CrossEntropyLoss	Function: create_model; classifier modified with nn.Sequential(Dropout, Linear)
EfficientNet_B0	5.3M	30 epochs	3	0.5	42	2	CrossEntropyLoss	Function: create_model; classifier: Linear → ReLU → Dropout → Linear
ResNet50	25.6M	30 epochs	3	0.5	42	2	CrossEntropyLoss	Function: create_model; fc replaced with nn.Sequential(Dropout, Linear)
VGG19	143.7M	30 epochs	3	0.5	42	2	CrossEntropyLoss	Function: create_model; classifier layer [6] replaced with nn.Linear
MobileNet_V2	3.5M	30 epochs	3	0.5	42	2	CrossEntropyLoss	Function: create_model; classifier[1] replaced with nn.Linear

3.4.4 - Max-Voting Approach

After obtaining predictions from each model, implement a Max-Voting mechanism for each test sample, then select the class that receives the highest number of votes from the individual models. This approach helps to enhance accuracy and robustness by leveraging the strengths of each model. The Max-Voting ensemble method is an effective and reliable approach for skin cancer classification. The use of pre-trained deep learning models contributes to accurate and prompt diagnosis, has the potential to improve patient outcomes and may

contribute to reducing mortality rates in future clinical applications, and shows significant promise for medical image analysis and additional diagnostic tasks.

Mechanism: In the Max-voting ensemble mechanism, each model outputs a class prediction per image, and the most frequent class among the votes is selected as the final decision. Hossain et al. [24] demonstrated that combining several state-of-the-art pre-trained deep learning models through max voting achieved 93.18% accuracy and an AUC of 0.932, confirming its effectiveness in skin cancer detection. While this approach enhances robustness by leveraging diverse architectures, the absence of explicit tie-breaking rules remains a limitation, and strategies such as weighted voting or confidence-score averaging are recommended to further improve reliability.

Advantages:

- Decision stability; mitigates individual model variance.
- Simple and transparent; no probability calibration needed.
- Resilient when models have different calibration/threshold behaviors.

Limitations:

- Ignores confidence magnitudes; all votes weigh equally.
- A recognized limitation of the Max-voting ensemble approach is the lack of an explicit tie-breaking mechanism when multiple classes receive equal votes. To overcome this issue, several strategies can be implemented, including weighted voting based on model accuracy, confidence-score averaging, or hierarchical prioritization of clinically critical classes. Addressing this limitation is essential to ensure robust and reliable classification outcomes, as emphasized in recent work by Kumar et al. [24].

Why not Soft-Voting: Soft-Voting relies on calibrated probabilities; miscalibration across heterogeneous models can bias the blend. Max-Voting provides robust, calibration-agnostic consensus suited to clinical decision framing.

Imbalance handling techniques:

Advanced techniques to resolve class imbalance between frequent and rare categories, ensuring full support for the detection development system are as follows. As depicted in Table (3.4):

Table (3.4): Advanced Techniques used in the proposed method.

Technique	Description	Academic Rationale
WeightedRandomSampler	A sampling strategy that balances class distribution by assigning higher sampling probability to rare classes.	Ensures that minority classes are adequately represented during training, mitigating imbalance and improving model sensitivity to rare but critical categories.
FocalLoss	A modified loss function: $FL(p_t) = -\alpha_t(1-p_t)^{\gamma} \log(p_t)$ It down-weights easy examples and focuses on hard, misclassified samples.	Addresses severe class imbalance by emphasizing difficult cases, enhancing detection of rare conditions such as melanoma in medical imaging.
Alpha_tensor	A normalized weighting scheme: $w_i^{(norm)} = \left(\frac{1}{freq_i}\right) \times \left[\frac{K}{\sum_{j=1}^K \left(\frac{1}{freq_j}\right)}\right]$ Assigns higher weights to rare classes based on inverse frequency.	Provides a mathematically grounded approach to class weighting, ensuring rare classes contribute proportionally to training without destabilizing optimization.

• **WeightedRandomSampler:**

Idea: Bias per-batch sampling toward rare classes using class weights, increasing their presence without synthetic data.

Why chosen:

Minimal code changes in DataLoaders; straightforward and effective.

Preserves original data distribution; maintains clinical realism.

Synergizes with FocalLoss by providing exposure while the loss stresses hard examples.

Reduces nv dominance; improves exposure to mel, akiec, df.

Implementation: Compute class weights from training distribution and pass to the sampler used by DataLoaders.

• **Focalloss:**

The general equation for Focal Loss is:

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \dots\dots\dots (3.1)$$

The general Formula of Focalloss given by Equation (3.1).

Where:

- p_t : The predicted probability for the true class, representing the model's confidence in correctly classifying a sample.
- α_t : A class-specific weighting factor that balances the importance of positive/negative examples or rare/common classes.
- γ (gamma): A focusing parameter that reduces the relative loss for well-classified examples, allowing the model to concentrate on hard or misclassified samples.
- $FL(p_t)$: The computed focal loss for a given sample.

Purpose: Focal Loss is designed to address class imbalance by down-weighting easy examples and focusing training on hard, misclassified cases. This is particularly valuable in medical imaging, where rare but critical conditions (e.g., melanoma) must be emphasized.

• **Alpha_tenser:**

The normalized weight for class i is defined as:

$$w_i^{(norm)} = \left(\frac{1}{freq_i} \right) \times \left[K / \sum_{j=1}^K \left(\frac{1}{freq_j} \right) \right] \dots\dots\dots (3.2)$$

The general Formula of Alpha_tensor given by Equation (3.2).

Where:

- $w_i^{(norm)}$: Normalized weight assigned to class i . Reflects the rarity of the class, giving higher weights to rare classes.
- $freq_i$: Frequency of class i , calculated as N_i / N_{total} , where N_i is the number of samples in class i and N_{total} is the total number of samples.
- K : Total number of classes in the dataset.
- $\sum_{j=1}^K \left(\frac{1}{freq_j} \right)$: Summation of inverse frequencies across all classes, ensuring normalization and stability of weights.

Purpose: Alpha_tensor provides a principled way to assign weights to classes based on their rarity, ensuring that rare classes are emphasized during training without excessively inflating their contribution.

• **Shared Objective:** Both Focal Loss and Alpha_tensor aim to mitigate the challenge of class imbalance:

- Focal Loss achieves this by adjusting the loss function to prioritize hard and rare examples.

- Alpha_tensor achieves this by scaling class weights inversely to their frequency, ensuring balanced representation.

Common Goal: Together, they enhance the sensitivity of deep learning models to rare but clinically significant categories, such as melanoma, while maintaining numerical stability and preventing overfitting to majority classes. This synergy is crucial in medical imaging tasks where diagnostic accuracy for minority classes directly impacts patient outcomes.

- **Training setup and Hyperparameters:**

- **Batch_size** = (16) Small batches help capture fine medical details, raise the chance of rare class presence per batch, and fit typical Kaggle GPU memory. As shown in Table (3.5).

- **Num_epochs:** In this study, the number of training epochs was set to 30, which falls within a balanced range of 25–40 epochs generally effective for datasets of comparable size. Training for fewer than 25 epochs risks under-learning rare classes due to insufficient exposure, while exceeding 40 epochs tends to promote overfitting, thereby reducing generalization performance. This observation aligns with empirical evidence in medical imaging research, where prolonged training has been shown to correlate positively with overfitting and negatively with generalization capacity [71].

- **Optimizer:** AdamW with moderate weight decay. AdamW, weight decay is decoupled (applied separately), which improves generalization, reduces overfitting and improving training stability. Unlike Adam optimizer, where weight reduction is linked to gradient updates.

AdamW Optimization Algorithm: AdamW serves as an extension of the Adam optimizer, introducing decoupled weight decay to bolster generalization and lessen overfitting in deep learning models. The algorithm operates through three essential phases: calculating moving averages of gradients, performing bias correction, and adjusting parameters using a decoupled weight decay factor. This will be elaborated upon in the subsequent steps:

1. Calculating Moving Averages of Gradients (Moments):

At each training step t , the first-order and second-order moment estimates are updated as follows:

- $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$.
- $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$.
- g_t : gradient of the loss function with respect to the parameter at step t .
- m_t : first-order moving average (analogous to momentum).
- v_t : second-order moving average (analogous to variance).

- β_1, β_2 : exponential decay rates for the moving averages (commonly $\beta_1 \approx 0.9$, $\beta_2 \approx 0.999$).

2. Performing Bias Correction:

To counteract initialization bias in the moving averages, bias-corrected estimates are computed:

- $\widehat{m}_t = \frac{m_t}{1 - \beta_1^t}$
- $\widehat{v}_t = \frac{v_t}{1 - \beta_2^t}$
 - $\widehat{m}_t$: bias-corrected first-order estimate.
 - $\widehat{v}_t$: bias-corrected second-order estimate.

3. Parameter Update Rule in AdamW:

The parameter update in AdamW is given by:

$$\bullet \theta_{t+1} = \theta_t - Lr \left(\frac{\widehat{m}_t}{\sqrt{\widehat{v}_t + \epsilon}} + \lambda \cdot \theta_t \right) \dots \dots \dots (3.3)$$

- θ_t : parameter (weight) at step t .
- Lr : learning rate.
- ϵ : small constant to prevent division by zero (typically 10^{-8}).
- λ : weight decay coefficient.
 - $\frac{\widehat{m}_t}{\sqrt{\widehat{v}_t + \epsilon}}$: standard Adam update term.
- $\lambda \cdot \theta_t$: decoupled weight decay term, unique to AdamW.
- **Regularization**: Dropout and label smoothing to reduce overconfidence and overfitting.
- **Scheduler for VGG19**:

In training the VGG19 model, the ReduceLROnPlateau learning rate scheduler was applied to stabilize optimization. VGG19 is a deep network with a heavy architecture and a large number of parameters. Since it lacks Batch Normalization layers, it often suffers from high gradient variance and sensitivity to learning rate changes. At the beginning of training, the model typically achieves relatively low validation accuracy (around 12–16%), which is noticeably weaker than more modern architectures such as EfficientNet-B0, MobileNetV2, and DenseNet121 even though DenseNet121 has 121 layers. To address this, the scheduler monitors the validation loss and automatically reduces the learning rate whenever progress stalls. This adjustment helps improve validation accuracy, maintain stable training accuracy, and reduce the risk of overfitting. As this model exhibits sensitivity to the learning rate, it is

modified using Focalloss with Smoothing. In addition, the learning rate is scheduled with a gamma value of 1.5 and a smoothing value of 0.05, as illustrated in the Table (3.6) that details the Training Settings for each of the five pre-trained models. This strategy is implemented to promote training, prevent fluctuations in val_loss, stabilize val_acc, ensure the models are trained correctly and consistently, and achieve the expected results.

Table (3.5): Training Settings and Hyperparameters.

Setting	Value	Academic Rationale
Batch Size	16 for each model	A moderate batch size balances computational efficiency with gradient stability, ensuring reliable convergence without overwhelming memory resources.
Num_Epochs	30 for each model	Provides sufficient iterations for the model to learn complex patterns while avoiding overfitting through early stopping.
Optimizer	AdamW	AdamW combines adaptive learning rates with weight decay, improving generalization compared to standard Adam. and reduces overfitting.
Learning Rate Scheduler	ReduceLROnPlateau	Dynamically reduces the learning rate when validation loss plateaus, enhancing convergence and preventing stagnation.
Weight Saving	Best weights saved	Preserves the most effective model state during training, ensuring optimal performance is retained.

Table (3.6): Details Training Settings for Each Pre-Trained Model.

Model	Gamma Value	Learning Rate	Weight Decay	Dropout	Alpha	Smoothing	Reduction
DenseNet121	2.0	0.0001	1e-4	0.3	alpha_tensor	—	Mean
EfficientNet-B0	2.0	0.0001	—	0.4	alpha_tensor	—	Mean
ResNet50	2.0	0.0001	1e-4	0.3	alpha_tensor	—	Mean
VGG19	1.5	0.0001	1e-4	0.5	alpha_tensor	0.05	Mean
MobileNetV2	2.0	0.0001	1e-4	0.3	alpha_tensor	—	Mean

- All models employed Focal Loss, integrating gamma and alpha parameters to boost sensitivity to infrequent classes like melanoma (mel), while lessening the effect of more prevalent classes such as nevus (nv).

- In the Focal Loss framework, the parameter gamma acts as a focusing factor that regulates the relative contribution of easy versus hard samples. While a value of 2.0 is commonly adopted to strengthen learning from misclassified cases without disregarding correctly classified ones, in the VGG19 model gamma was reduced to 1.5 to moderate this emphasis, thereby avoiding excessive penalization of easy samples and ensuring more stable convergence with consistent validation performance. Moreover, VGG19 utilized a modified version of Focal Loss with the smoothing parameter set at 0.05, which further enhanced training stability by mitigating overconfidence in predictions.

- All models featured Dropout techniques intended to mitigate overfitting and boost generalization performance.

- The ReduceLROnPlateau scheduler was employed in the training of VGG19 to dynamically adjust the learning rate, thereby stabilizing the optimization process and mitigating fluctuations in validation loss and accuracy. Configured with mode='min', factor=0.5, and patience=3, the scheduler reduced the learning rate by half whenever no improvement in validation loss was observed across three consecutive epochs. This adaptive mechanism enhanced convergence stability and prevented the optimizer from stagnating in suboptimal minima, ultimately facilitating the achievement of the desired diagnostic performance.

- The application of a reduced learning rate contributed to the stabilization of training and yielded precise and satisfactory training results.

3.4.5 - Evaluation metrics

Evaluation metrics play a crucial role in evaluating the effectiveness of machine learning models, especially in medical imaging applications where the precision of diagnoses significantly affects patient outcomes. In multi-class skin lesion classification, metrics must capture not only overall correctness but also sensitivity to minority classes, balance between false positives and false negatives, and the ability to diagnose misclassification patterns. The following subsections present the key evaluation metrics Precision, Recall, F1-score, Overall Accuracy and Confusion Matrix.

3.4.5.1 Accuracy:

Accuracy is a fundamental metric used to evaluate the performance of classification models. It measures the proportion of correct classifications relative to the total number of instances in the dataset. Formally, accuracy can be expressed as:

$$\text{Accuracy} = \frac{\text{Correct classifications}}{\text{Total classifications}} = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots (3.4)$$

where TP (True Positives) represents the number of positive instances correctly predicted, TN (True Negatives) represents correctly predicted negative instances, FP (False Positives) represents the number of negative instances incorrectly classified as positive, and FN (False Negatives) represents the number of positive instances incorrectly classified as negative.

Usage: Provides a global measure of performance. Effective when classes are balanced, but misleading under severe imbalance (e.g., NV dominating HAM10000).

3.4.5.2 Recall:

Recall, also known as sensitivity or true positive rate, assesses the model's capability to identify all relevant positive instances. It is defined as:

$$\text{Recall} = \frac{\text{Correctly classified actual positives}}{\text{All actual positives}} = \frac{TP}{TP+FN} \dots\dots\dots (3.5)$$

This metric is particularly significant in scenarios where false negatives carry a high cost, highlighting the importance of capturing as many true positive instances as possible.

Usage: Measures purity of predictions. Critical when false positives are costly, such as misclassifying benign lesions as malignant.

3.4.5.3 Precision:

Precision quantifies the accuracy of the positive predictions made by the model. It is calculated using the following formula:

$$\text{Precision} = \frac{\text{Correctly classified actual positives}}{\text{Everything classified as positive}} = \frac{TP}{TP+FP} \dots\dots\dots (3.6)$$

Usage: Measures purity of predictions. Critical when false positives are costly, such as misclassifying benign lesions as malignant.

3.4.5.4 F1-score:

The F1-score combines precision and recall into a single metric, providing a balance between the two by taking their harmonic mean. It is expressed as:

$$F1\text{-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \dots\dots\dots (3.7)$$

Usage: Harmonic mean balancing Precision and Recall. Suitable under imbalance where Accuracy alone is insufficient.

Confusion matrix: Visualize misclassifications, especially *nv* vs *mel*.

Usage: Diagnoses misclassification patterns, e.g., confusion between *mel* and *bnl*. Provides basis for deriving TP, FP, FN, TN per class.

In multi-class classification settings, metrics such as Precision, Recall, and F1-Score must be reported beyond binary formulations. Specifically, per-class values should be presented to highlight performance disparities across categories, while macro-averaging ensures equal weighting of all classes and weighted-averaging reflects the true distribution of the dataset. This combined reporting strategy provides a balanced and clinically meaningful evaluation, particularly for underrepresented but critical classes such as melanoma.

3.4.5.5 Macro-Averaged F1:

Formula:

$$\text{Macro-F1} = \frac{1}{N} \sum_i^N F1_i \dots\dots\dots (3.8)$$

Where:

- N : The total number of classes in the multi-class classification task.
- $F1_i$: The F1-Score calculated for class i .
- $\sum_i^N = \sum F1_i$: The summation of F1-Scores across all classes.
- $\frac{1}{N}$: Arithmetic mean, ensuring each class contributes equally regardless of its size.

3.4.5.6. Weighted-Averaged F1:

Formula:

$$\text{Weighted-F1} = \frac{\sum_{i=1}^N n_i \cdot F1_i}{\sum_{i=1}^N n_i} \dots\dots\dots (3.9)$$

Where:

- N : The total number of classes in the multi-class classification task.

- $F1_i$: The F1-Score calculated for class i .
- n_i : The number of samples belonging to class i .
- $\sum_i^N n_i$: The total number of samples across all classes.
- $\sum_i^N n_i \cdot F1_i$: Weighted summation of F1-Scores, where each class's contribution is proportional to its sample size.

Table (3.7): Comprehensive Overview of Evaluation Metric Formulations and Their Analytical Contexts in Binary and Multi-Class Classification.

Metric Type	Mathematical Expression	Analytical Description	Applicability and Context
Recall (Sensitivity)	$\frac{TP}{TP + FN}$	Quantifies the model's ability to correctly identify positive instances among all actual positives. High recall is essential when false negatives carry significant clinical risk.	Critical in medical diagnostics, e.g., detecting melanoma cases.
Precision (Positive Predictive Value)	$\frac{TP}{TP + FP}$	Measures the proportion of correctly predicted positives among all predicted positives. High precision minimizes false alarms.	Important when false positives are costly, such as misclassifying benign lesions as malignant.
F1-Score	$\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$	Harmonic mean balancing precision and recall; provides a unified measure of classification performance.	Suitable for evaluating overall model robustness.
Macro-Averaged F1	$\frac{1}{N} \sum_i^N F1_i$	Averages F1-Scores across all classes equally, ensuring minority classes contribute equally to the overall metric.	Ideal for imbalanced datasets with clinically rare categories.
Weighted-Averaged F1	$\frac{\sum_i^N n_i \cdot F1_i}{\sum_i^N n_i}$	Weights each class's F1-Score by its sample	Appropriate for real-world datasets

		size, reflecting the true data distribution.	with uneven class frequencies.
Micro-Averaged F1	Computed globally from aggregated TP, FP, FN across all classes	Evaluates overall performance by treating all classes collectively; emphasizes dominant categories.	Suitable for large, balanced datasets.

This table (3.7) presents a detailed synthesis of evaluation metric formulations and their interpretative significance in both binary and multi-class classification frameworks. It delineates how each metric quantifies model performance, emphasizing the importance of macro and weighted averaging for achieving balanced and clinically meaningful assessments, particularly in imbalanced medical datasets such as dermoscopic lesion classification.

The proposed methodology represents a coherent integration of thorough data preparation and sophisticated modeling strategies. The process commenced with a stratified data split to guarantee balanced representation across all classes, followed by systematic preprocessing procedures. Architectural diversity was achieved through the implementation of five pre-trained models, which were later combined using a Max-Voting ensemble approach to yield reliable and accurate predictive outcomes. To address the issue of class imbalance, techniques such as WeightedRandomSampler, Focal Loss, and Alpha_tensor were utilized. These methods improved the model's sensitivity to rare classes while mitigating the influence of majority categories, thus promoting more robust and generalizable performance. As depicted in Table (3.8).

The training configuration was purposefully selected to guarantee robust and consistent learning. Specifically, a small batch size (16), an adequate number of epochs (30), and the AdamW optimizer with multiple regularization techniques were employed.

Learning-rate scheduling for VGG19 was implemented through the ReduceLROnPlateau scheduler, while the best model weights were preserved and early stopping was applied. Collectively, these measures minimized overfitting, stabilized the training process, and systematically enhanced validation accuracy (val_acc) by reducing validation loss (val_loss).

Table (3.8): Presenting a Summary of the Proposed Methodology.

Component	Description	Academic Rationale
Dataset	HAM10000 dermoscopic image dataset with seven skin lesion categories.	Provides a globally recognized benchmark with clinical diversity, ensuring comparability with international studies.
Preprocessing	Image resizing, normalization, and stratified splitting into train/val/test.	Ensures data quality, balanced representation, and reproducibility across experiments.
Selection of pre-trained models	DenseNet121, EfficientNet-B0, ResNet50, VGG19, MobileNetV2.	Captures complementary feature representations through diverse CNN architectures, enhancing robustness.
Ensemble Strategy(Max-Voting Approach)	Max-Voting fusion of model outputs.	Improves decision stability and reliability by combining multiple models.
Imbalance Handling	WeightedRandomSampler, FocalLoss, Alpha_tensor.	Addresses class imbalance by emphasizing rare but clinically significant categories.
Training Settings and Hyperparameters.	Batch size=16, epochs=30, AdamW optimizer, ReduceLROnPlateau scheduler, best weight saving.	Provides controlled optimization, prevents overfitting, and ensures convergence.
Evaluation Metrics	Precision, Recall, F1-score, Overall Accuracy, Macro Averaged F1, Weighted Averaged F1, and Confusion Matrix.	Offers comprehensive performance assessment, highlighting sensitivity to rare classes.

CHAPTER 4:

RESULTS AND DISCUSSION

4.1 Introduction

This chapter offers a thorough analysis of the experimental results obtained from the proposed deep learning framework focused on the classification of dermoscopic skin lesions. The evaluation is designed to underscore both the quantitative performance metrics and the qualitative implications of the methodology utilized. The results are examined in the context of previously identified challenges, which include substantial intra-class variability, inter-class similarity, and a significant class imbalance found in the HAM10000 dataset. An evaluation of the performance of five pre-trained convolutional neural networks DenseNet121, EfficientNet-B0, ResNet50, VGG19, and MobileNetV2 was conducted under standardized training conditions, utilizing advanced methodologies such as WeightedRandomSampler, Focal Loss with alpha weighting, and for VGG19, a ReduceLROnPlateau scheduler to facilitate stable learning. These techniques were essential in mitigating the prevalence of common classes (e.g., nevus) while increasing sensitivity to rare but clinically important categories like melanoma. The discussion integrates the reported accuracy of 88.92% and a macro-averaged F1 score of 0.8492. Furthermore, the computational efficiency of the framework requiring approximately 4 hours and 55 minutes of training with moderate GPU memory usage (3–6 GB per model) is analyzed in the context of reproducibility and sustainability in resource-constrained environments. In addition, comparisons were conducted with four recent studies M. M. Hossain et al. [24], A. Rahman [22], L. Wang et al. [29], and R. Kaur et al. [67] to evaluate the suggested methodology in relation to current methods. The results of these comparative analyses will be discussed later in this chapter to emphasize the comparative advantages and drawbacks of the proposed framework.

4.2 Observation and Elaboration

In this section, a summary of the main findings extracted from the experimental results of the recommended framework is presented. The analysis emphasizes performance patterns across different models, concentrating on accuracy, sensitivity, and computational efficiency. Each result is detailed to elucidate its significance in tackling dataset imbalance and clinical relevance. These findings lay the foundation for an in-depth discussion of the enhancements realized and comparative insights.

4.2.1 Results of Accuracy of Individual Models

In Figure (4.1), a detailed comparative analysis of five pre-trained convolutional neural networks (CNNs) DenseNet121, EfficientNet-B0, ResNet50, VGG19, and MobileNetV2 is

provided under identical experimental conditions using the HAM10000 dataset. All models were trained with consistent preprocessing, sampling, and loss strategies to ensure methodological rigor and comparability. The observed accuracies ranged between 83.57% and 87.97%, reflecting relatively close performance across the architectures. ResNet50 achieved an accuracy of 87.97%, while DenseNet121 followed with 87.87%, both benefiting from advanced architectural designs that promote efficient gradient flow and feature reuse. MobileNetV2 attained 87.22%, demonstrating that lightweight architectures can deliver competitive results while maintaining computational efficiency. EfficientNet-B0 reported 86.67%, consistent with its compound scaling approach, though showing slight sensitivity to minority classes. VGG19, in contrast, recorded the lowest accuracy of 83.57%, which can be attributed to its conventional deep stack of convolutional layers lacking residual connections or dynamic scaling. This design makes it more vulnerable to fluctuations in learning rates and less adaptable to class imbalance, while its comparatively heavy classifier head requires extended training to stabilize, collectively constraining its convergence efficiency. Despite these observations, the differences in accuracy remain modest. Without the inclusion of confidence intervals or statistical significance testing, it is not possible to assert with certainty that one model is definitively superior to the others. Accordingly, these findings should be interpreted as indicative trends rather than conclusive evidence of comparative advantage. This limitation underscores the necessity of complementing accuracy comparisons with robust statistical validation to substantiate claims of performance differences.

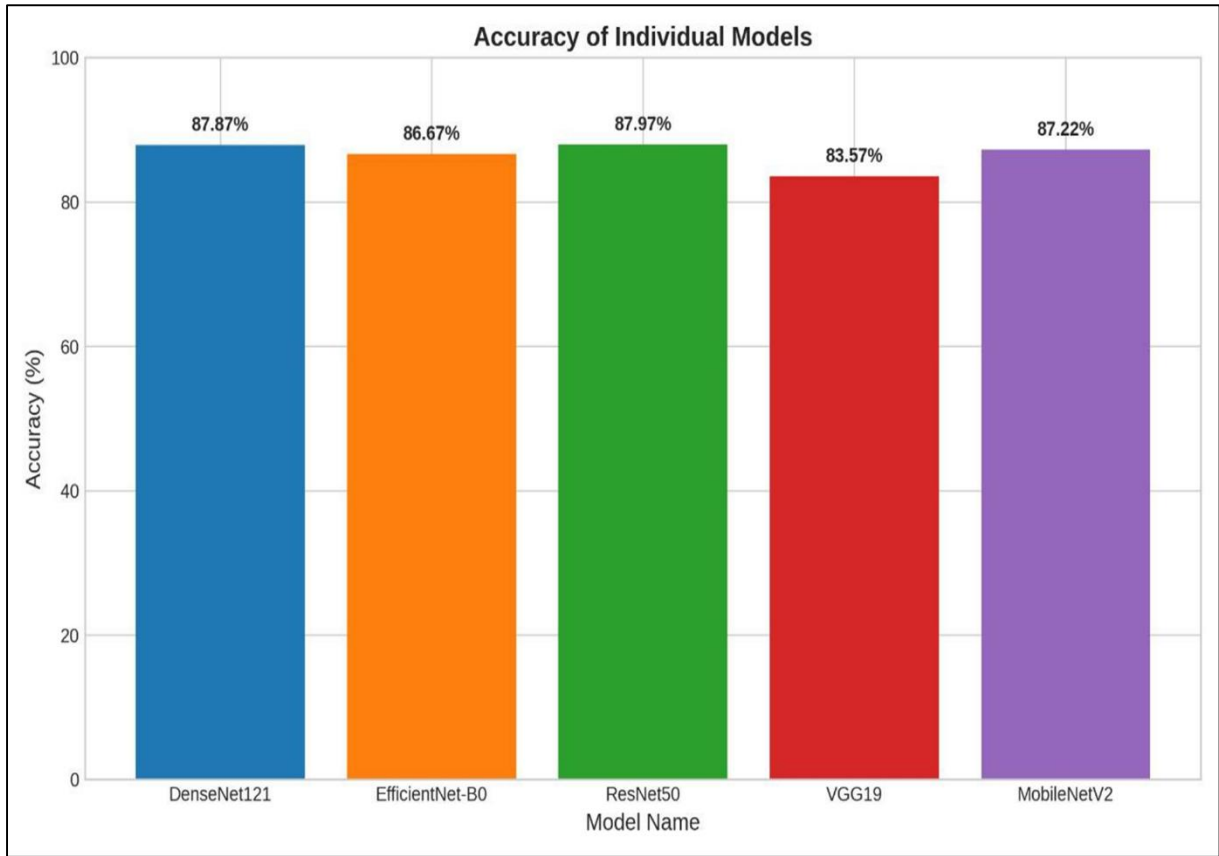


Figure (4.1): The Results of Accuracy of Individual Models.

4.2.2 Results of F1-Score of Individual Models

Figure (4.2) presents a thorough comparison of the F1-scores achieved by five different pre-trained convolutional neural networks (CNNs) utilized in the proposed framework: DenseNet121, EfficientNet-B0, ResNet50, VGG19, and MobileNetV2. These models were trained under consistent conditions using the HAM10000 dataset, with uniform preprocessing, sampling, and loss strategies to ensure a fair assessment. The F1-score, characterized as the harmonic mean of precision and recall, acts as a balanced measure for evaluating classification performance, especially in contexts where class imbalance is present. Among the five models, ResNet50 attained the highest F1-score of 0.8786, demonstrating its remarkable ability to maintain both sensitivity and specificity across different lesion categories. DenseNet121 closely trailed with a score of 0.8774, leveraging its densely connected architecture that improves feature reuse and gradient flow. MobileNetV2 attained a competitive score of 0.8734, affirming its efficiency and adaptability despite its lightweight structure. EfficientNet-B0 produced an F1-score of 0.8658, which is slightly lower than the leading models, yet it still exhibits stable classification performance due to its compound scaling approach.

In comparison, VGG19 achieved the lowest F1-Score of 0.8369, underscoring its limited capacity to maintain a balance between precision and recall when subjected to the same training conditions. This finding can be ascribed to its traditional architecture, which is devoid of residual connections and dynamic scaling mechanisms, thus making it less robust to class imbalance and more reactive to fluctuations in learning rates. Additionally, VGG19's comprehensive classifier head requires extended training times to reach stability, and its deeper convolutional framework might struggle to adapt efficiently to the diverse lesion patterns in the dataset. These structural constraints collectively led to a diminished performance in comparison to more sophisticated and optimized models. The differences in F1-scores across the models underscore the critical significance of architectural design and training dynamics in the classification of medical images. These findings support the ensemble strategy utilized in the proposed method, which leverages the complementary advantages of individual models to achieve a more balanced and resilient classification result.

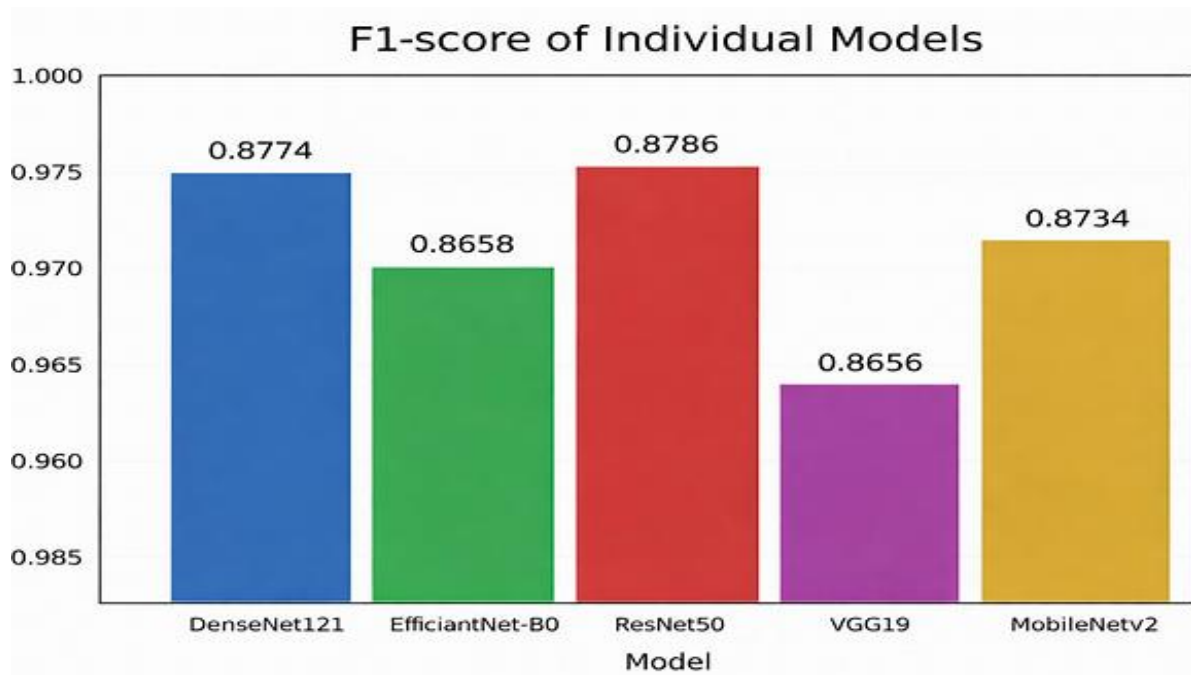


Figure (4,2): Presents The Results of F1-Score of Individual Models.

4.2.3 Results of Per-Class F1-score Comparison Across Disease Categories

Figure (4.3) provides a comprehensive analysis of the F1-scores for each class attained by the various pre-trained models within the suggested framework, emphasizing the classification efficacy across seven unique skin lesion categories: akiec, bcc, bkl, df, mel, nv, and vasc. These scores illustrate the model's capability to maintain a balance between precision and recall for every class, thereby presenting a detailed perspective on its diagnostic

dependability. The nevus (nv) category achieved the highest F1-score of 0.94, reflecting both its prevalence in the dataset and the model's ability to integrate its characteristics. The nv (nevus) category reached the highest F1-score of 0.94, which is consistent with its frequency in the dataset and the model's ability to integrate its features. In a comparable manner, the df (dermatofibroma) and vasc (vascular lesions) categories obtained impressive scores of 0.93 and 0.89 respectively, signifying the model's capability in identifying clearly defined and less ambiguous patterns. The benign keratosis-like lesions (bkl) category obtained a score of 0.79, a result that can be associated with its heterogeneous appearance and the overlap it shares with both benign and malignant categories, making accurate classification challenging. The bcc (basal cell carcinoma) category achieved an F1-score of 0.87, indicating dependable performance even with its moderate presence in the dataset. Notably, the melanoma (mel) category achieved the lowest F1-score of 0.70, which constitutes a critical limitation in the system's diagnostic performance. This outcome reflects the considerable intra-class variability of melanoma and its visual similarity to other pigmented lesions, particularly within imbalanced datasets. Despite the application of focal loss and various sampling strategies, sensitivity to melanoma remained restricted, thereby exposing a prominent weakness in the models. Consequently, the Melanoma result must be emphasized as a key limitation, underscoring the need for further methodological improvements such as targeted augmentation or specialized subnetworks to enhance detection of this clinically significant category.

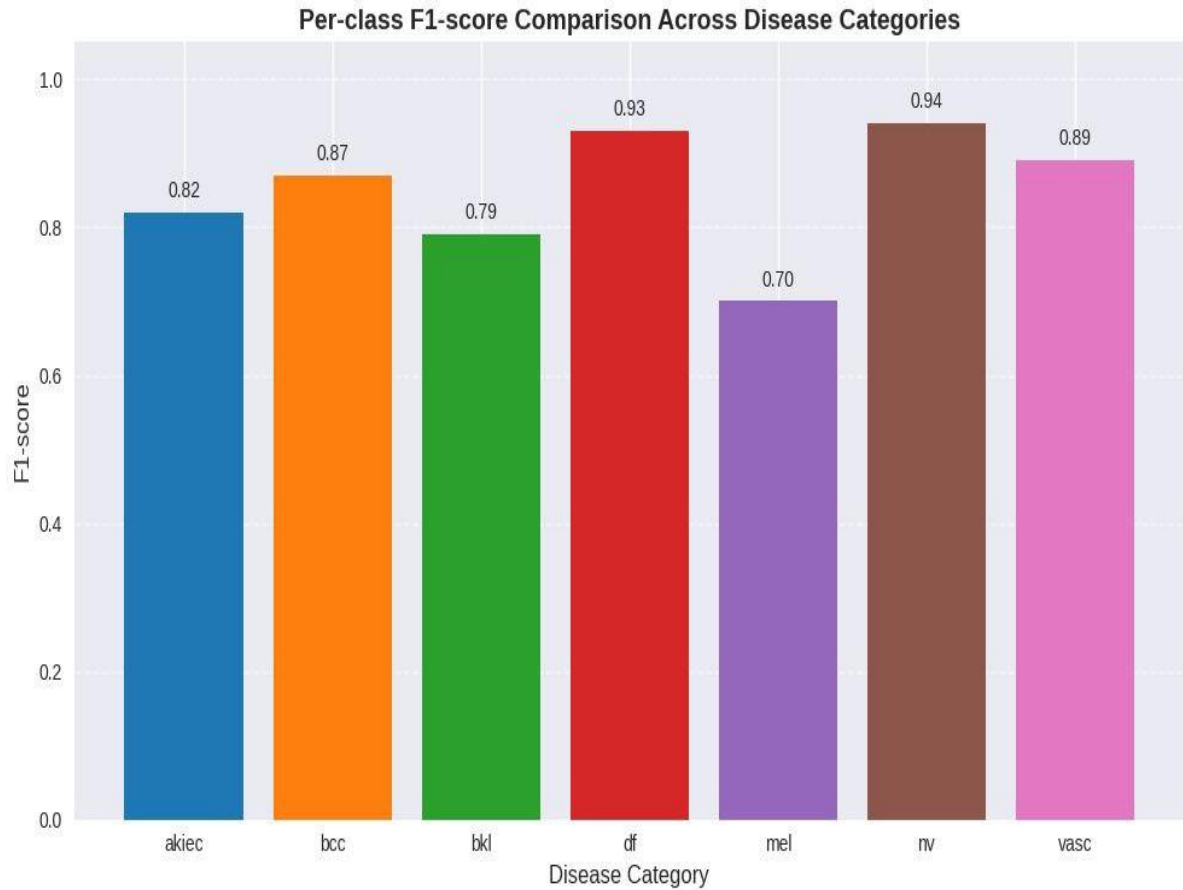


Figure (4.3): The Results of Per-Class F1-score Comparison Across Disease Categories.

In summary, the analysis conducted on a per-class basis demonstrates that although the proposed framework effectively addresses dominant and structurally unique classes, it encounters challenges in identifying rare and visually intricate lesions, such as melanoma. These findings underscore the importance of performing class-specific assessments in the realm of medical image classification and advocate for the use of ensemble methods and sophisticated loss functions to improve sensitivity towards minority classes.

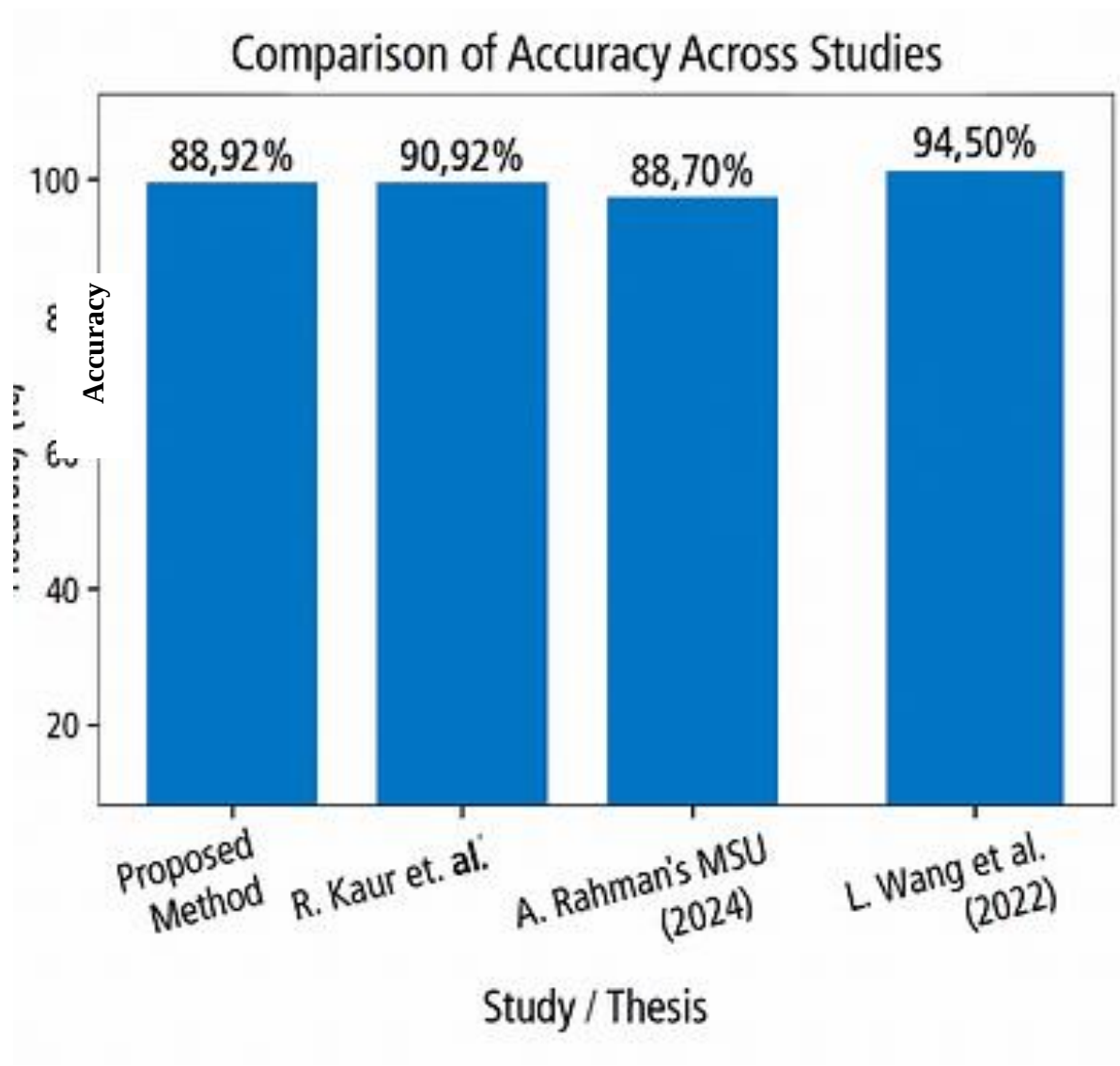


Figure (4.4): Comparative Analysis of Classification Accuracy Among Recent Studies and the Proposed Method.

4.2.4 Discussions

Figure (4.4) displays a vertical bar chart that showcases the classification accuracy (%) attained by five different methods for skin lesion classification, one of which is the Proposed Method. The x-axis lists the various methods, while the y-axis represents the accuracy percentage. Each bar reflects the performance of a specific model in skin lesion classification and shows the following values. As depicted in Table (4.1).

Table (4.1): Presents The Comparative Analysis of Classification Accuracy Values Among Recent Studies and the Proposed Method.

Study / Method	Accuracy (%)
Proposed Method	88.92%
R. Kaur et al.	90.92%
A. Rahman’s MSU	88.70%
M. M. Hossain et al	93.18%
L. Wang et al.	94.50%

To compare the proposed method with the recent studies mentioned in the Table (4.1), in terms of strengths and weaknesses or shortcomings between the proposed method and these studies, as follows:

– **Strengths of the Proposed Method**

- **Transparency and Reproducibility:** The methodology delivers an extensive overview of execution time (4h 55m) and the environment (Kaggle Notebook Editor), thereby reinforcing academic rigor and reproducibility.

- **Multi-DCNN Architecture:** The configuration of multiple fine-tuned pre-trained CNNs affirms the model’s durability and diversity.

- **Transparency and Reproducibility:** The methodology provides a clear account of execution time (4h 55m) and the computational environment (Kaggle Notebook Editor), thereby reinforcing academic rigor and reproducibility. In addition, performance efficiency was comprehensively documented through both execution time and memory usage, ensuring precise evaluation of resource demands. Particular attention was devoted to addressing class imbalance and improving classification accuracy in datasets characterized by high visual similarity among lesions. This was achieved through strategies that enhanced sensitivity to minority classes and strengthened diagnostic reliability. Finally, the framework was validated through seven-class prediction with performance evaluation metrics including precision, recall, F1-score, and overall accuracy providing a transparent and rigorous assessment of the system’s effectiveness.

- **Educational Accessibility:** This framework is intended to operate within open settings, making it particularly suitable for educational and instructional applications.

– Weaknesses Compared to Other Studies

- **Reduced Accuracy:** The Proposed Method attained an accuracy of 88.92%, which is inferior to all other studies referenced.
- **Prolonged Execution Time:** The method demands 4 hours and 55 minutes, thus it is markedly slower compared to R. Kaur’s model, which achieves the task in only 76 seconds.
- **Absence of Pre-trained or Transformer Models:** Differing from the strategies used by Wang and Hossain, the Proposed Method fails to include advanced architectural frameworks.

In summary, Figure (4.4) distinctly demonstrates the comparative performance of the Proposed Method in light of contemporary research. Although it does not surpass the other models regarding raw accuracy, its advantages in transparency, reproducibility, and educational significance render it a noteworthy academic contribution. Future enhancements may aim at incorporating advanced architectures and refining execution time to improve clinical relevance.

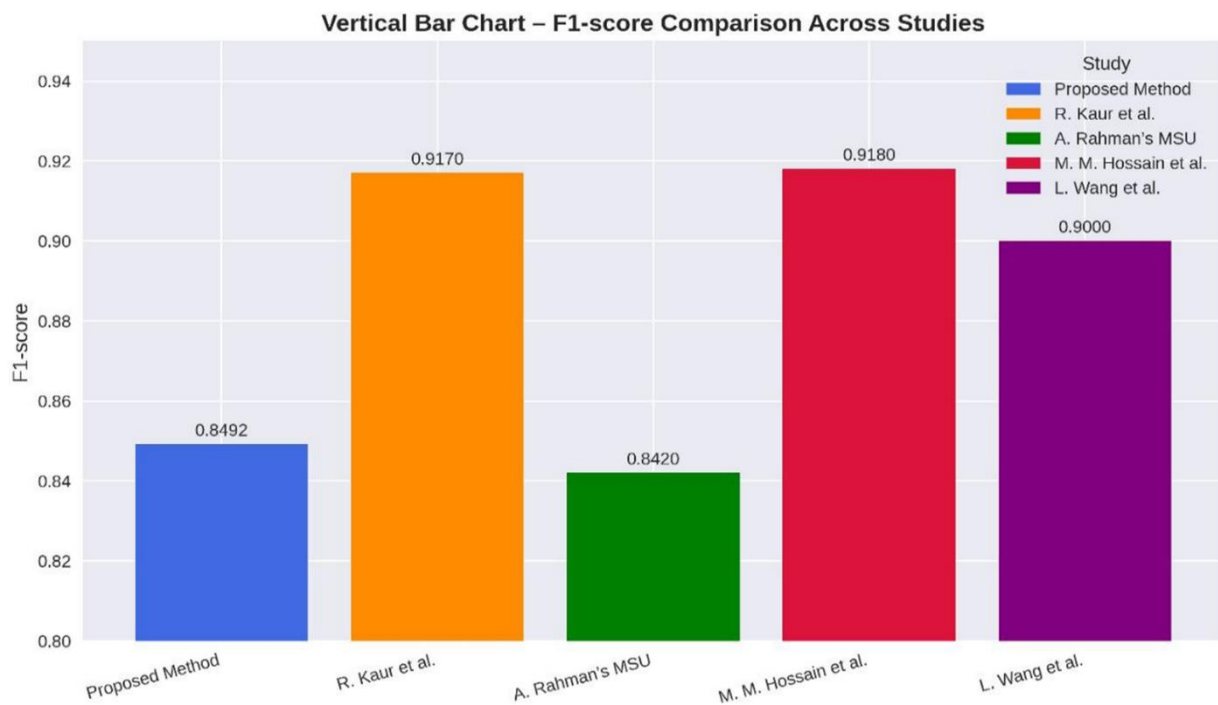


Figure (4.5): F1-score Comparison Among Different Skin Lesion Classification Approaches.

In Figure (4.5), a vertical bar chart illustrates the F1-scores obtained through five distinct methods of skin lesion classification. The x-axis features the names of the studies, while the y-axis denotes the F1-score values. For further numerical specifics, refer to Table (4.2).

Table (4.2): Presents The Comparative Analysis of Classification F1-Score Values Among Recent Studies and the Proposed Method.

Study / Method	F1-score
Proposed Method	0.8492
R. Kaur et al.	0.9170
A. Rahman's MSU	0.8420
M. M. Hossain et al.	0.9180
L. Wang et al.	0.9000

To compare the proposed method with the recent studies mentioned in the Table (4.2), in terms of strengths and weaknesses or shortcomings between the proposed method and these studies, as follows:

- The Proposed Method achieved an F1-score of 0.8492, which is lower than most recent studies.
- R. Kaur et al. [67] and M. M. Hossain et al. [24] demonstrated the strongest performance (0.9170 and 0.9180 respectively). However, the proposed method demonstrated its efficiency with a significantly shorter execution time of 4 hours and 55 minutes compared to the study by M. M. Hossain et al. [24], which required 12 hours of execution, despite the considerable similarity in terms of the Kaggle operating environment. Both studies also utilized multi-DCNN and the Max-Voting strategy, as shown in Table (4.3).
- A. Rahman's MSU. [22] demonstrated performance comparable to the Proposed Method (0.8420). A. Rahman's thesis achieved slightly lower performance than the proposed method and was almost identical (0.8420), but the proposed method had the advantage of being fully documented for the execution time as shown in Table (4.3). In addition, this proposed method provided a comprehensive analysis of the F1- scale results for each of the seven categories of skin lesions: akiec, bcc, bkl, df, mel, nv, and vasc. This is what distinguishes this proposed method from all the recent studies that were compared to it, as shown in Table (4.3).

Table (4.3): Presents Comprehensive Classification Metrics and Max-Voting Ensemble Evaluation for Seven Skin Lesion Categories Using the Proposed Method.

Class	Precision	Recall	F1-score	Support
Akiec	0.78	0.86	0.82	65
Bcc	0.88	0.85	0.87	103
Bkl	0.77	0.82	0.79	220
Df	0.95	0.91	0.93	23
Mel	0.71	0.70	0.70	223
Nv	0.94	0.94	0.94	1341
Vasc	0.92	0.86	0.89	28
Overall Accuracy			0.89	2003
Macro Average	0.85	0.85	0.85	2003
Weighted Average	0.89	0.89	0.89	2003

- L. Wang et al. secured a distinguished F1-score of 0.9000, aligning with methods that are based on Vision Transformers.

In summary, Figure (4.5) and Table (4.2) collectively depict the comparative efficacy of the Proposed Method in relation to recent research. Although the Proposed Method is academically sound and thoroughly documented, prospective enhancements could aim at improving classification performance to meet or surpass the standards established by ensemble and transformer-based models.

Table (4.4): Presents a Summary of Comparison of Skin Lesion Classification Studies and Proposed Method.

Study / Method	Approach	Year	Dataset	Accuracy	F1-Score	AUC	Execution Time and memory usage	Environment
Proposed Method	Multi-DCNN Max-Voting Approach	2026	HAM10000	88.92%	0.8492	–	4h 55m 3–6 GB per model.	Kaggle Notebook Editor. GPU T4 x 2
R. Kaur et al.	N-DCNN Novel Deep Convolutional Neural Network	2025	ISIC 2020	90.92%	0.9170	–	76 seconds	Device/GeForce GTX 1080 Ti
A. Rahman's MSU	EfficientNet-B0 (Transfer Learning)	2024	HAM10000	88.70%	0.8420	–	Not specified	Python + TensorFlow/Keras
M. M. Hossain et al.	Pre-trained DL Models + Max Voting	2024	ISIC 2018 + HAM10000	93.18%	0.9180	0.932	12 hours	Kaggle Notebook Editor. GPU T4x2, P100 and TPU: VM v3-8
L. Wang et al.	Vision Transformers (ViTs)	2022	ISIC 2018 + HAM10000	94–95%	0.9000	0.932	Not specified	Research environment (GPU required)

Table (4.5): Methodological Comparison of Skin Lesion Classification Studies.

Study / Method	Dataset Scope	Number of Classes	Split Protocol	External Validation	Computational Environment
Proposed Method (2026)	HAM10000	7 classes	Step 1: 80% train / 20% test. Step 2: train split into 80% final train / 20% validation → Final: 64% train, 16% validation, 20% test	No external dataset used	Kaggle Notebook Editor, GPU T4 × 2, 3–6 GB per model
R. Kaur et al. (2025)	ISIC 2020	2 classes (MEL vs BEN)	70% train / 10% validation / 20% test	No external validation	GeForce GTX 1080 Ti
A. Rahman’s MSU (2024)	HAM10000	7 classes	Not specified	No external validation	Python + TensorFlow/Keras
M. M. Hossain et al. (2024)	ISIC 2018 + HAM10000	7 classes	80% train / 10% validation / 10% test	Partial (HAM10000 cross-dataset validation)	Kaggle Notebook Editor, GPU T4 × 2, P100, TPU VM v3-8
L. Wang et al. (2024)	ISIC 2018 + HAM10000	7 classes	70% train / 15% validation / 15% test	Yes (tested on ISIC 2020)	Research environment, GPU A100

This table (4.5) presents a structured comparison of methodological frameworks used in recent skin lesion classification research, including the proposed method. It outlines differences in dataset scope, number of classes, data-splitting protocols, validation strategies, and computational environments factors that significantly affect the fairness and interpretability of reported outcomes.

• **Dataset Scope and Number of Classes:** The proposed method utilizes the HAM10000 dataset, which contains seven lesion categories, enabling a multiclass classification task that mirrors real-world diagnostic complexity. In contrast, Kaur et al. [67] used ISIC 2020, a binary dataset (melanoma vs. benign), representing a simpler classification problem. Studies such as Hossain et al. [24] and Wang et al. [29] combined ISIC 2018 and HAM10000, maintaining seven classes and providing a more balanced multiclass evaluation.

• **Split Protocols:** The proposed method applies a two-stage data-splitting process: first dividing the dataset into 80 % training and 20 % testing, then splitting the training subset into 80 % final training and 20 % validation. This yields final proportions of 64 % training, 16 % validation, and 20 % testing, ensuring balanced learning and reliable evaluation. Other studies adopted different ratios Wang et al. [29] used 70/15/15, Hossain et al [24]. used 80/10/10, and Kaur et al. [67] used 70/10/20 each influencing generalization and performance consistency.

• **Validation Approach:** The proposed method relies on internal validation, as all subsets originate from the same dataset (HAM10000). This ensures consistency within the same data distribution but does not evaluate generalization to unseen external data. In contrast, Wang et al. [29] performed external validation using ISIC 2020, providing stronger evidence of robustness. Hossain et al. [24] conducted partial cross-dataset validation, while other studies limited their evaluation to internal splits.

• **Computational Environment:** The proposed method was implemented in the Kaggle Notebook Editor, using dual GPU T4 units with memory usage between 3 GB and 6 GB per model. This setup offers a practical and reproducible environment but is less powerful than the A100 GPU used by Wang et al. [29] Hardware variations from GTX 1080 Ti to TPU VM v3-8 affect training time, scalability, and achievable precision.

The table summarizes how variations in dataset composition, number of classes, and validation strategies affect the comparability of skin lesion classification results. The proposed method demonstrates consistency and methodological precision through structured internal validation and balanced data partitioning within the HAM10000 dataset, ensuring reliable outcomes. In conclusion, incorporating external validation using independent datasets such as ISIC 2018 or ISIC 2020 would further enhance generalization capability and reinforce the credibility of future comparative studies.

4.2.4.1 DenseNet121 Results

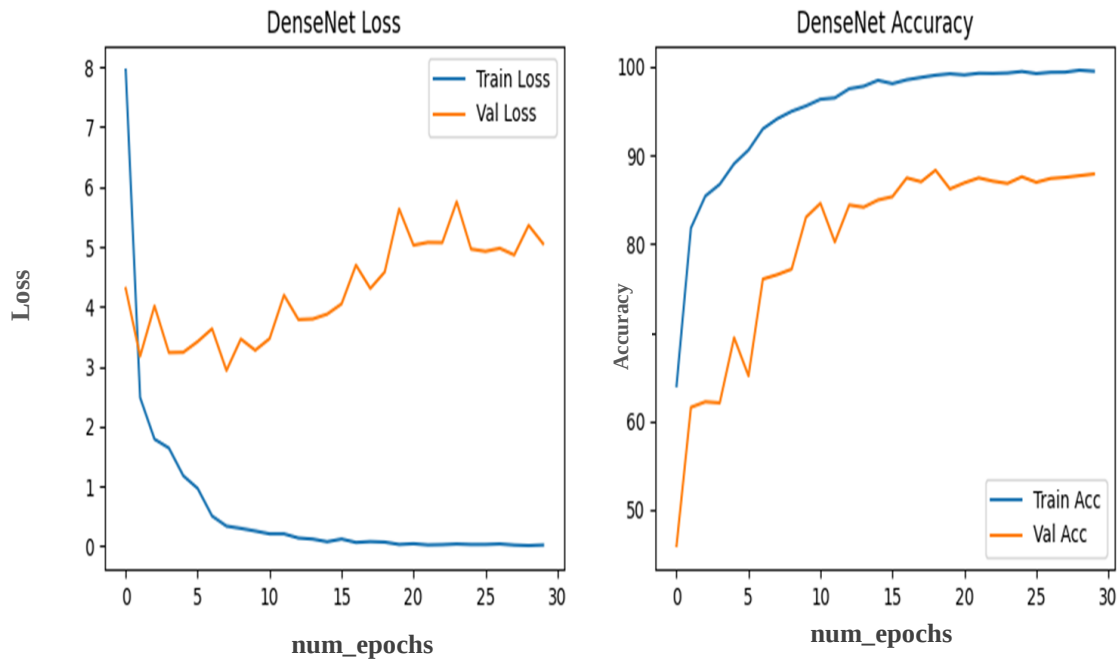


Figure (4.6): Detailed Analysis of DenseNet121 Training and Validation Performance Over 30 Epochs.

In Figure (4.6), the performance of the DenseNet121 model was evaluated across 30 epochs, providing insights into both convergence and generalization. The training process exhibited a consistent decline in loss, decreasing from 7.9497 at the initial epoch to 0.0208 at the final epoch. In parallel, training accuracy improved substantially from 64.02% to 99.50%, reflecting strong convergence and effective learning of the training data. Validation performance followed a different trajectory. Validation accuracy rose markedly from 45.98% in the first epoch to 87.90% in the final epoch, indicating that the model was able to generalize to unseen samples. However, validation loss fluctuated throughout the epochs, beginning at 4.3025 and ending at 5.0547. This variability is consistent with the use of Focal Loss, which is designed to emphasize minority classes by penalizing misclassifications more heavily than CrossEntropyLoss. Unlike CrossEntropyLoss, which typically produces lower and more stable loss values due to its bias toward frequent classes, Focal Loss introduces higher initial loss values and greater fluctuations, reflecting its sensitivity to rare but clinically significant categories. An important observation is the disparity between training and validation performance. While training accuracy reached near perfect levels ($\approx 99\%$) from epoch 19 onwards, validation accuracy stabilized around 87%. This gap suggests the presence of partial overfitting, rather than a fully well regularized learning process. The model demonstrates robustness in capturing minority classes, yet the divergence between training and validation

metrics highlights the need for cautious interpretation. Overall, Figure (4.7) illustrates both the strengths and limitations of the proposed DenseNet121 architecture. The model achieves high accuracy and sensitivity to rare lesion categories, which is particularly valuable in medical imaging tasks characterized by class imbalance. Nevertheless, the observed fluctuations in validation loss and the performance gap between training and validation accuracy underscore the importance of further measures such as external validation, statistical significance testing, and additional regularization techniques to confirm the model’s generalizability and mitigate overfitting risks.

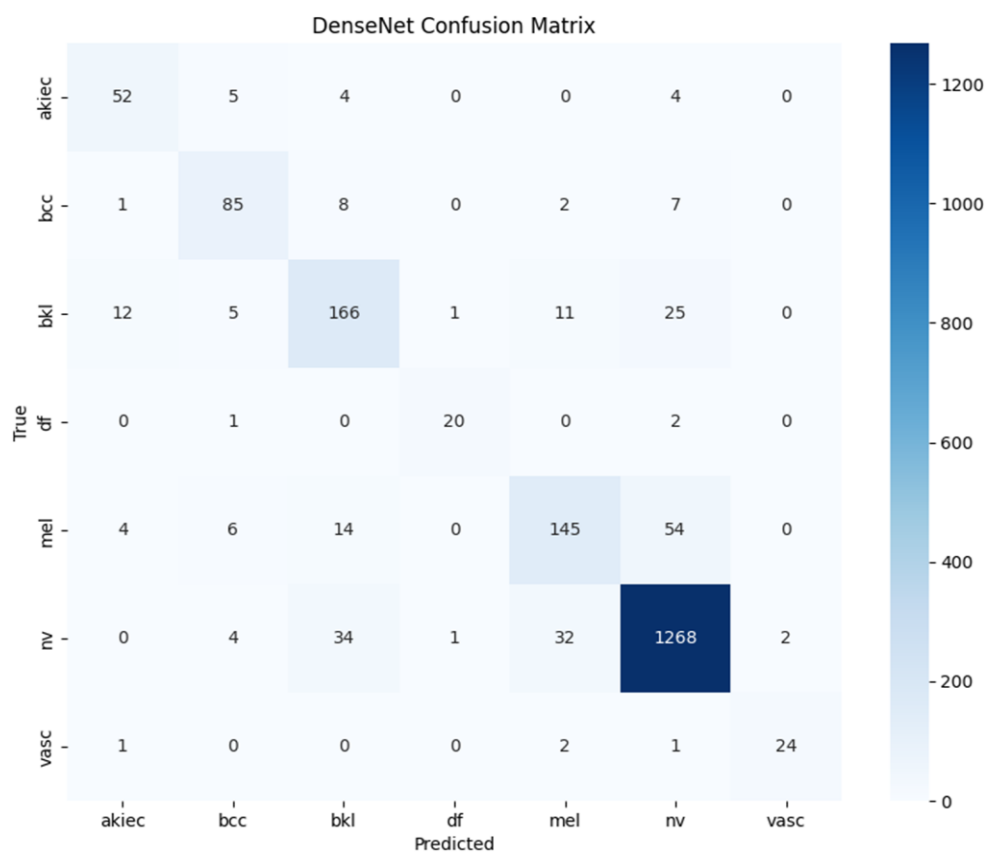


Figure (4.7): Presents DenseNet121-confusion-matrix Analysis.

Figure (4.7) shows the confusion matrix for the DenseNet121 model, which assesses its classification performance across seven categories of skin lesions: akiec, bcc, bkl, df, mel, nv, and vasc. The diagonal cells depict correct predictions, with the 'nv' class reaching the highest accuracy (1268 correct out of 1341), emphasizing its relevance in the dataset. The model exhibits strong performance on 'df' (20/23) and 'vasc' (24/28), indicating its proficiency in effectively recognizing rare classes. Instances of misclassification have been identified in

'mel', with 54 cases wrongly categorized as 'nv', and in 'bkl', which was inaccurately classified as both 'nv' and 'mel' (25 and 11 occurrences, respectively). The 'akiec' category demonstrates moderate performance (52/65), with some confusion identified between 'bcc' and 'bkl'. In summary, the matrix indicates that although DenseNet121 achieves high accuracy for both prevalent and rare classes, it still faces difficulties with inter-class overlap, especially among pigmented lesions like bkl and mel. The gradient of colors represents the density of predictions, where darker shades signify a higher frequency of correct classifications.

4.2.4.2 *EfficientNet-B0 Results*

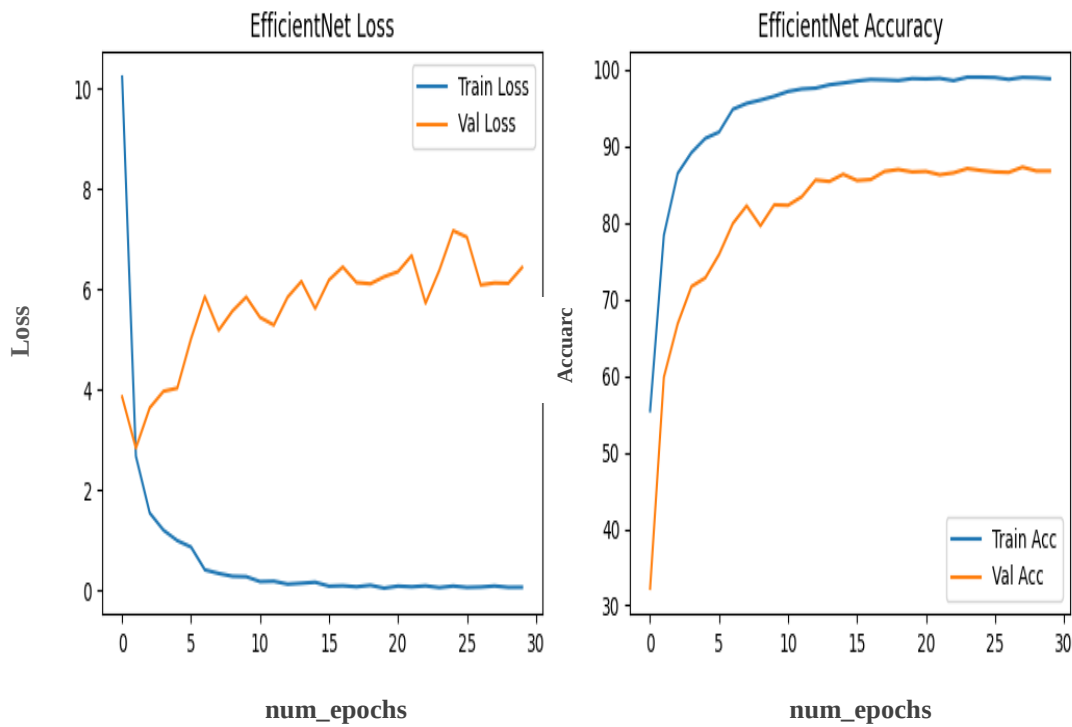


Figure (4.8): Detailed Analysis of EfficientNet-B0 Training and Validation Performance Over 30 Epochs.

In Figure (4.8), the EfficientNet-B0 model demonstrates rapid convergence and effective feature extraction across 30 epochs. During the first epoch, training loss is high (10.22) with modest accuracy (55.47%), while validation accuracy is low (32.25%). By the fifth epoch, training loss falls below 1.0 and accuracy exceeds 91%, confirming efficient optimization. From the sixth epoch onward, training accuracy steadily rises to 99.03%, with loss stabilizing near zero, indicating strong convergence. Validation accuracy shows a consistent upward trajectory, reaching 87.27% by epoch 28, whereas validation loss fluctuates

between 3.8 and 7.1. These fluctuations are expected when employing Focal Loss, which emphasizes minority classes and imposes heavier penalties on their misclassification compared to CrossEntropyLoss. This dynamic reflects the model’s attempt to balance learning across both majority and minority categories. However, the disparity between near-perfect training accuracy ($\approx 99\%$) and lower validation accuracy ($\approx 87\%$) requires cautious interpretation. While the rising validation accuracy suggests that the model generalizes reasonably well, the persistent gap between training and validation metrics may indicate partial overfitting, rather than a fully well-regularized learning process. Thus, the results highlight both the strengths and limitations of EfficientNet-B0: it achieves high accuracy and sensitivity to rare classes, but further measures such as external validation, statistical testing, and additional regularization are necessary to confirm robustness and mitigate overfitting risks.

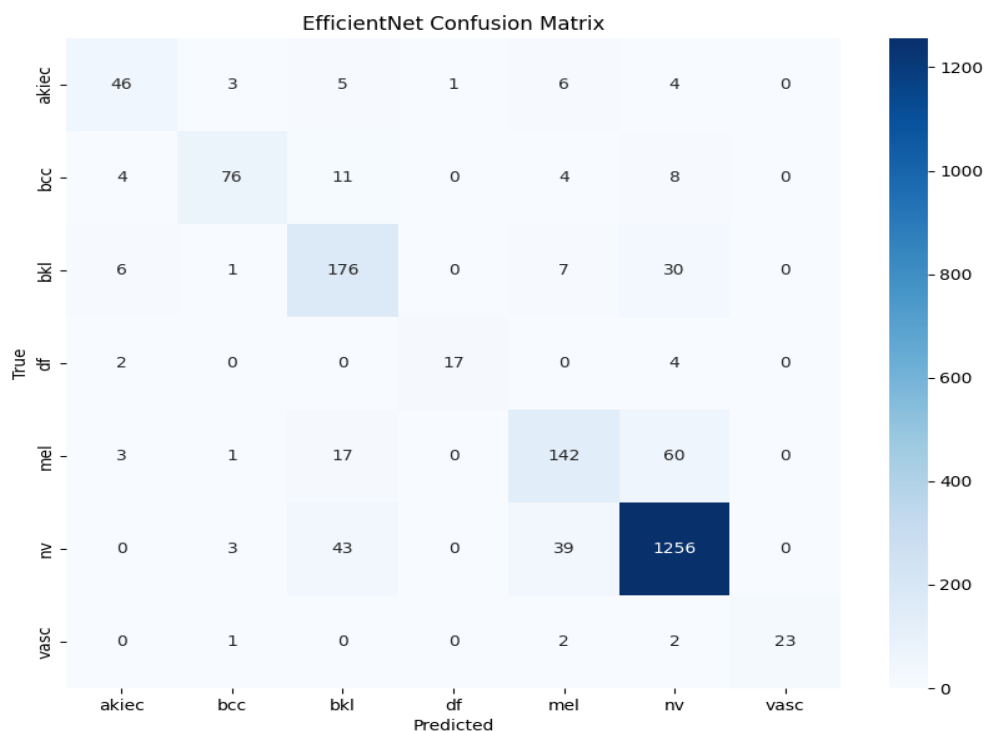


Figure (4.9): Presents EfficientNet-B0 Confusion-Matrix Analysis.

In Figure (4.9), the confusion matrix provides a visual representation of the classification efficacy of the EfficientNet-B0 model across seven distinct skin lesion categories: akiec, bcc, bkl, df, mel, nv, and vasc. The model demonstrates a high level of predictive accuracy for the predominant class nv, achieving 1256 correct predictions with

minimal instances of misclassification, which indicates a bias towards categories that are well-represented. In contrast, the rare classes *df* and *vasc* show a lack of classification accuracy, with only 17 and 23 correct predictions, respectively. This phenomenon is characteristic of models that are trained on imbalanced datasets, where classes that are underrepresented receive insufficient focus during the optimization process. The *mel* category, which is of significant clinical importance, records 142 correct predictions but is often misidentified as *nv* and *bkl*. Such misclassifications frequently occur in models that rely solely on visual features, highlighting the necessity for enhanced differentiation strategies.

In summary, although the model demonstrates robust performance on prevalent classes, it requires improvements in its handling of minority categories. The application of methods like Focal Loss or specific oversampling could enhance sensitivity to these infrequent classes without compromising overall accuracy.

4.2.4.3 ResNet50 Results

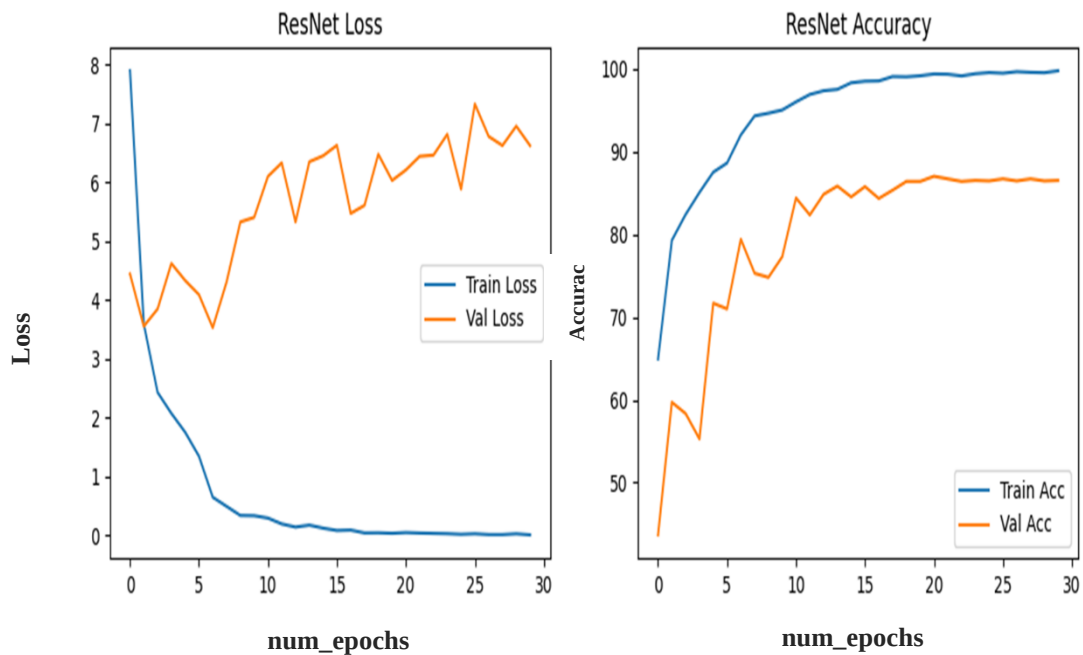


Figure (4.10): Detailed Analysis of ResNet50 Training and Validation Performance Over 30 Epochs.

In Figure (4.10), the ResNet50 model demonstrates rapid convergence and effective feature learning across 30 epochs. Training loss decreases sharply from 7.89 to nearly zero, while training accuracy rises from 64.94% to 99.78%, confirming strong optimization and reliable learning of the training data. Validation accuracy also shows a consistent upward

trajectory, improving from 43.67% to 86.53%, which indicates that the model is capable of generalizing to unseen samples.

However, validation loss begins at 4.44 and increases steadily, reaching 6.62 by the final epoch. This upward trend is partly attributable to the use of Focal Loss, which emphasizes minority classes and imposes heavier penalties on their misclassification compared to CrossEntropyLoss. While this explains the elevated and fluctuating loss values, the divergence between near-perfect training accuracy and lower validation accuracy requires careful interpretation. It is not sufficient to dismiss this gap as a “deliberate learning approach.” Instead, the combination of near-zero training loss and rising validation loss may suggest partial overfitting, even though validation accuracy continues to improve. This pattern highlights both the strengths and limitations of ResNet50: the model achieves high accuracy and sensitivity to rare classes, but the discrepancy between training and validation metrics underscores the need for further regularization, statistical testing, and external validation to confirm robustness and mitigate overfitting risks.

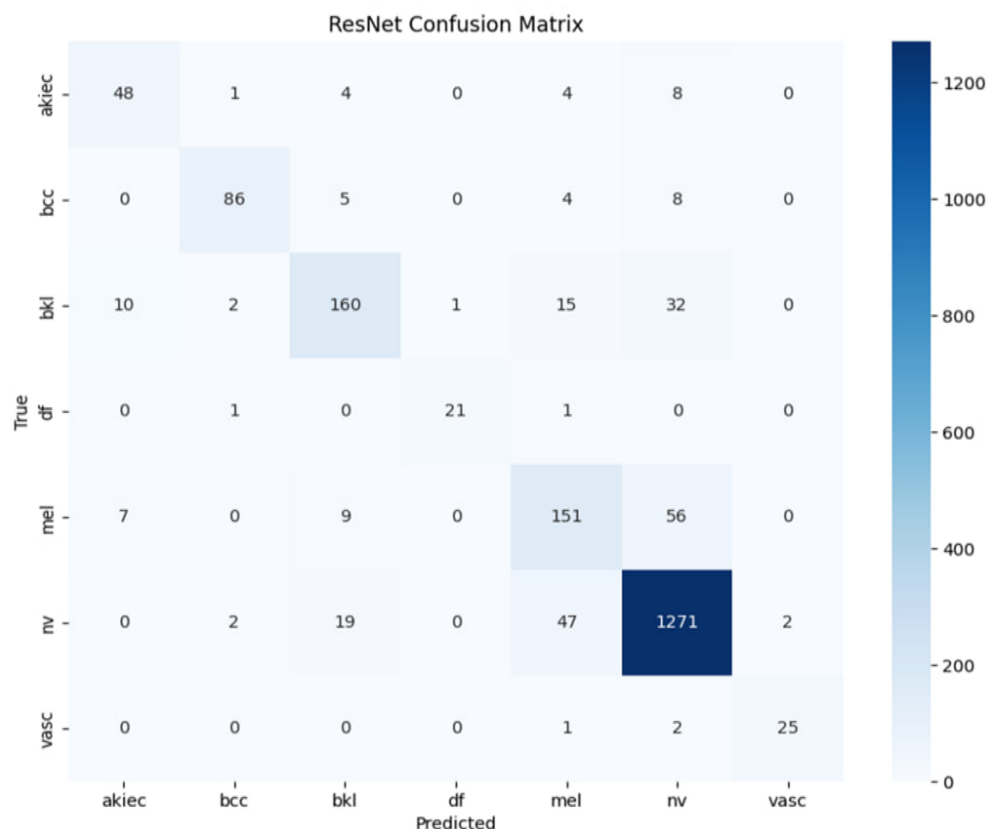


Figure (4.11): Presents ResNet50 Confusion-Matrix Analysis.

In Figure (4.11), the confusion matrix for ResNet50 reveals a solid classification performance for the leading class nv, attaining 1271 precise predictions and displaying limited misclassifications. This outcome underscores the model's inclination towards categories that are well-represented. In contrast, the rare classes df and vasc demonstrate limited recognition, with only 21 and 25 correct predictions respectively, which is a typical issue encountered in imbalanced datasets. The mel class, recognized for its clinical importance, has documented 151 accurate predictions; however, it is frequently misclassified as nv and bkl, underscoring the visual resemblances and the need for enhanced feature differentiation. The akiec and bcc classes demonstrate a moderate level of performance, with certain cases of misclassification occurring into nv and bkl, which suggests the existence of overlapping characteristics. In summary, the model showcases high precision for prevalent classes while facing challenges with minority categories. This disparity can be remedied by employing techniques like Focal Loss or class-aware sampling, which improve sensitivity to infrequent lesions while maintaining overall accuracy.

4.2.4.4 VGG19 Results

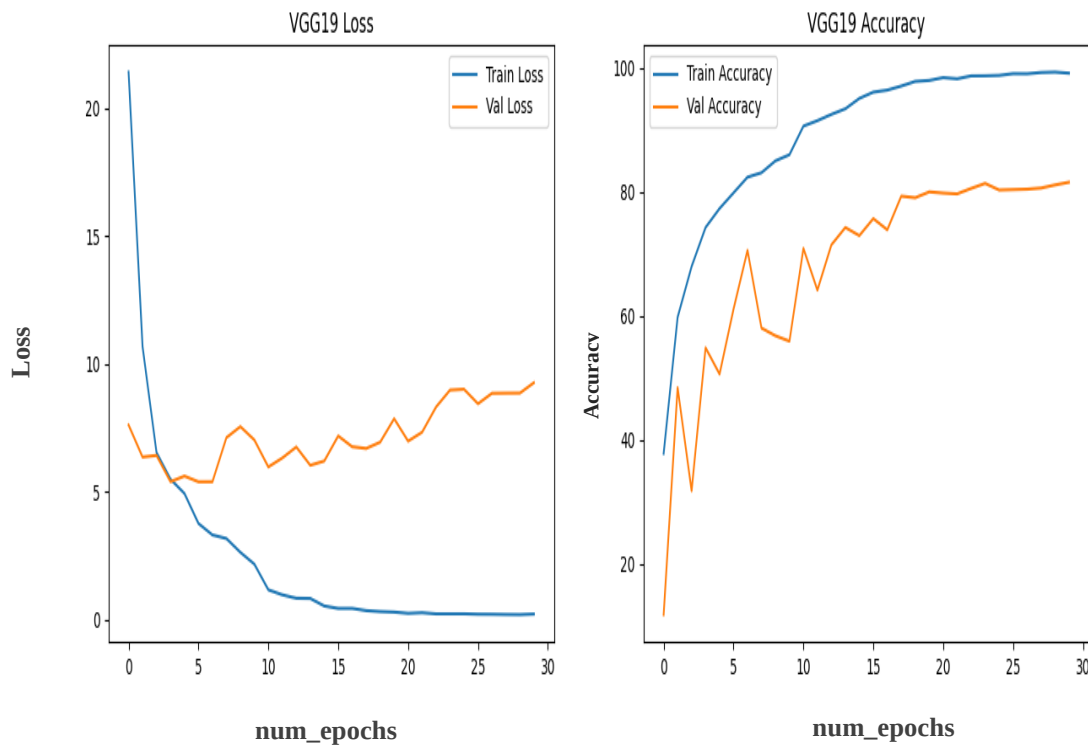


Figure (4.12): Detailed Analysis of VGG19 Training and Validation Performance Over 30 Epochs.

In Figure (4.12), the VGG19 model demonstrates a progressive learning curve across 30 epochs. Training loss decreases markedly from 21.42 to 0.21, while training accuracy rises from 37.85% to 99.19%, confirming strong convergence. Validation accuracy also improves steadily, increasing from 11.85% to 81.60%, which indicates that the model is capable of generalizing to unseen data.

However, validation loss fluctuates between 7.61 and 9.26 throughout the epochs. While such fluctuations are partly attributable to the use of Focal Loss, which emphasizes minority classes and penalizes their misclassification more heavily than majority classes, the persistent gap between near-zero training loss and relatively high validation loss requires cautious interpretation. It is not sufficient to categorically dismiss this behavior as a deliberate learning strategy; rather, the combination of very low training loss and elevated validation loss may suggest partial overfitting, even though validation accuracy continues to improve. The architectural characteristics of VGG19 further contribute to this instability. Unlike more modern architectures, VGG19 lacks Batch Normalization, making it more sensitive to learning rate fluctuations and prone to instability. Its deep layered structure amplifies this sensitivity, necessitating the use of a learning rate scheduler to stabilize training dynamics, track validation loss, and adapt optimization accordingly. Effective scheduling is therefore essential to mitigate overfitting risks and ensure consistent accuracy improvements, particularly in imbalanced datasets where the performance on rare classes is of critical clinical importance.

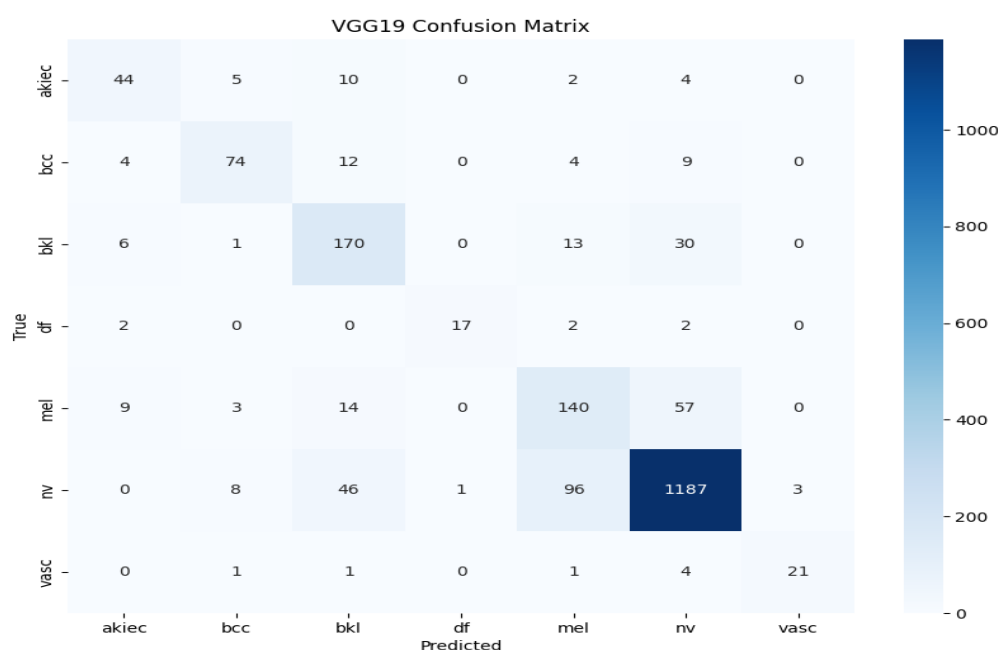


Figure (4.13): Presents VGG19 Confusion-Matrix Analysis.

In Figure (4.13), the confusion matrix of VGG19 shows a robust classification performance for the predominant class nv, achieving 1187 accurate predictions with only a few misclassifications. This result accentuates the model's bias towards categories that are significantly represented. In contrast, classes that are infrequent, like df and vasc, exhibit restricted identification, with only 17 and 21 correct predictions respectively, thereby emphasizing the challenges related to learning from imbalanced datasets. The mel class, recognized for its clinical importance, has documented 140 accurate predictions; however, it is frequently confused with nv and bkl, indicating a visual overlap and the need for enhanced feature differentiation. The misclassifications observed among akiec, bcc, and bkl indicate shared morphological characteristics that hinder effective boundary learning. To summarize, the model highlights strong precision for prevalent classes, while it exhibits lower sensitivity for minority categories. This discrepancy can be rectified by utilizing Focal Loss and class-aware sampling strategies, which enhance the detection of rare lesions without affecting overall accuracy.

4.2.4.5 MobileNetV2 Results

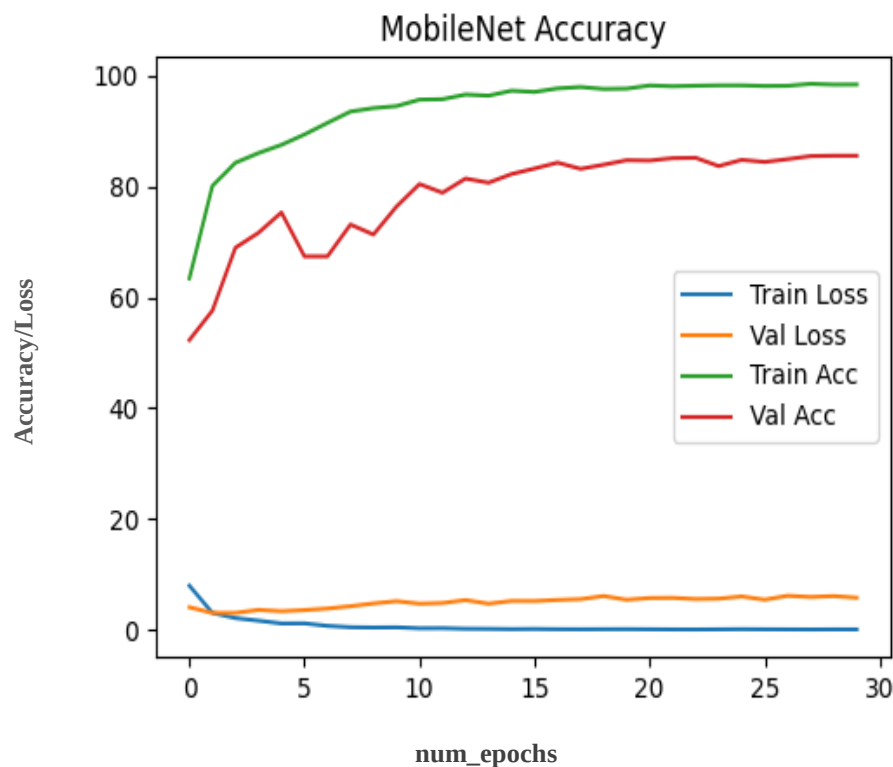


Figure (4.14): Detailed Analysis of MobileNetV2 Training and Validation Performance Over 30 Epochs.

In Figure (4.14), the MobileNetV2 model demonstrates effective and stable learning across 30 epochs. Training loss decreases from 7.76 to 0.059, while training accuracy rises from 63.29% to 98.49%, confirming strong convergence and efficient feature extraction. Validation accuracy also improves consistently, increasing from 51.53% to 85.78%, which indicates that the model is capable of generalizing across diverse lesion categories. Validation loss, however, fluctuates between 3.6938 and 5.7118. While such variation is partly attributable to the use of Focal Loss, which emphasizes minority classes and penalizes their misclassification more heavily than majority classes, the persistent gap between near-zero training loss and relatively elevated validation loss requires careful interpretation. It is not sufficient to categorically dismiss this behavior as a controlled learning strategy; rather, the combination of very low training loss and higher validation loss may suggest partial overfitting, even though validation accuracy continues to improve. The architectural design of MobileNetV2 contributes positively to its performance. Its lightweight structure and use of depthwise separable convolutions reduce computational complexity and help mitigate overfitting compared to deeper, more complex networks. Nevertheless, the observed discrepancy between training and validation metrics underscores the importance of further measures such as external validation, statistical testing, and potentially adaptive learning rate scheduling to confirm robustness and ensure consistent generalization.

Overall, Figure (4.14) highlights both the strengths and limitations of MobileNetV2: it achieves high accuracy and efficiency suitable for mobile-oriented medical applications, yet the validation loss dynamics necessitate cautious interpretation and additional safeguards to fully establish reliability in imbalanced medical datasets.

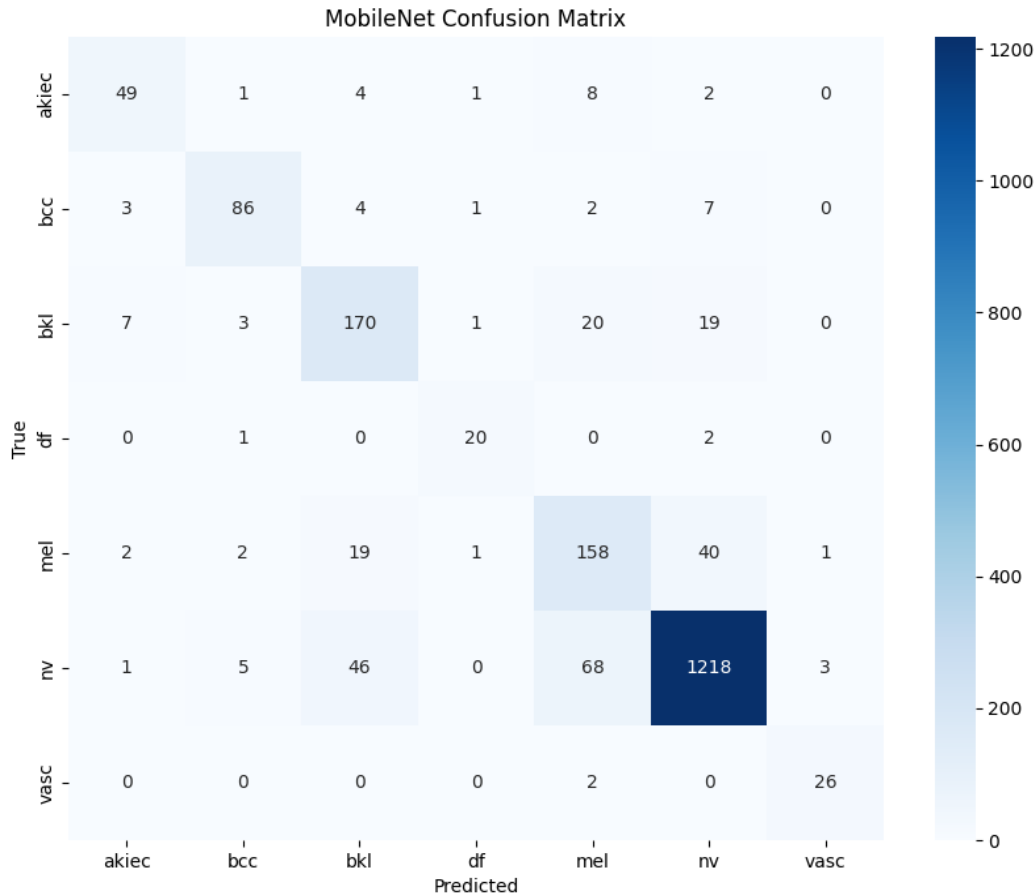


Figure (4.15): Presents MobileNetV2 Confusion-Matrix Analysis.

In Figure (4.15), the confusion matrix for MobileNetV2 illustrates robust classification capabilities across seven categories of skin lesions. The model demonstrates high precision for the predominant class *nv*, achieving 1218 correct predictions with minimal misclassifications, which indicates its proficiency in generalizing on prevalent patterns. In contrast, the rare classes *df* and *vasc* exhibit limited recognition, with only 20 and 26 correct predictions respectively, underscoring the difficulties associated with learning from imbalanced datasets. The clinically significant *mel* class records 158 correct predictions but is often confused with *nv* and *bkl*, suggesting visual similarities and the necessity for improved feature differentiation. The misclassifications observed among *akiec*, *bcc*, and *bkl* indicate shared morphological characteristics that hinder effective boundary learning. Nevertheless, the model sustains a balanced sensitivity and specificity, supported by the implementation of Focal Loss, which emphasizes the detection of rare classes.

In summary, the lightweight design of MobileNetV2, combined with its depthwise separable convolutions, improves its reliable performance, making it suitable for real-time medical applications without requiring adjustments to the learning rate.

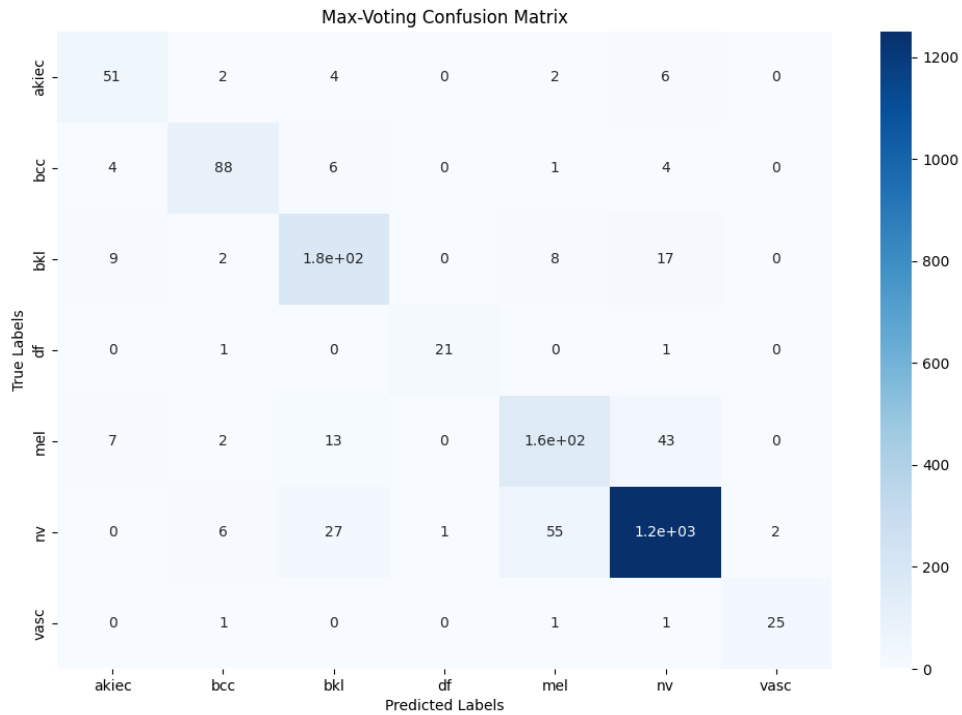


Figure (4.16): Presents Max-Voting Confusion-Matrix Analysis.

In Figure (4.16), the Max-Voting confusion matrix illustrates the classification performance of the ensemble model across seven categories of skin lesions. The nv class demonstrates strong predictive capability, with over 1200 correct predictions, reflecting high agreement among base models for common lesion patterns. In contrast, rare classes such as df and vasc achieve only 21 and 25 correct predictions, respectively, underscoring the ensemble's difficulty in capturing minority characteristics. The mel class, which is clinically significant, is correctly predicted in 160 cases; however, frequent misclassifications into nv and bkl highlight visual similarities and potential bias within the ensemble. Misclassifications among akiec, bcc, and bkl further suggest overlapping morphological traits that complicate decision boundaries. The Max-Voting strategy nonetheless enhances robustness by consolidating predictions, thereby reducing the variance of individual models and improving overall reliability. While the ensemble demonstrates improved sensitivity to rare lesions when combined with Focal Loss, the results must be interpreted with caution. The observed misclassifications and the absence of external validation or visual interpretability methods (e.g., Grad-CAM) limit the strength of claims regarding clinical applicability. Thus, the findings highlight the ensemble's potential for handling imbalanced datasets, but further validation and explainability analyses are required before considering deployment in clinical settings.

CHAPTER 5:
CONCLUSION AND FUTURE WORKS

5.1 Conclusion

This thesis introduced a structured deep learning framework for dermoscopic skin lesion classification, with a focus on ensemble learning. The pipeline encompassed dataset acquisition, preprocessing, augmentation, model selection, ensemble fusion, and performance evaluation, each stage designed to address challenges such as intra-class variability, inter-class similarity, and class imbalance. The HAM10000 dataset, comprising over 10,000 dermoscopic images across seven clinically significant categories (including melanoma, benign nevi, and keratoses), served as the primary benchmark. Its diversity in lesion morphology, image quality, and lighting conditions provided a robust foundation for evaluating the proposed framework and ensured clinical relevance in dermatological image classification. To maintain consistency and improve model performance, a series of preprocessing steps were implemented, including (Resizing, Horizontal Flip Rotation, ToTensor, Normalize), and employing stratified splitting to maintain class proportions across training, validation, and test sets, stratified splitting was performed at the image-level rather than the patient-level. A major challenge in this study was the class imbalance in the HAM10000 dataset, where benign nevi dominate while melanoma cases are scarce. To mitigate this issue, three complementary techniques were applied: WeightedRandomSampler to balance sampling, Focal Loss to emphasize minority classes, and Alpha-tensor weighting to adjust class importance, thereby improving model robustness and reliability. Along with carefully fine-tuned training parameters (batch size=16, epochs=30, AdamW optimizer, ReduceLROnPlateau scheduler, and optimal weight saving). The WeightedRandomSampler assured that all mini-batches during training maintained a balanced representation of classes by assigning increased sampling probabilities to categories that were underrepresented. This approach inhibited the model from displaying bias in favor of majority classes and improved its capacity to learn from infrequent instances. This thesis developed a deep learning framework for dermoscopic skin lesion classification, integrating imbalance-handling strategies such as Focal Loss, WeightedRandomSampler, and Alpha_tensor weighting, which collectively enhanced sensitivity to rare lesions like melanoma. Five pre-trained CNN architectures (DenseNet121, EfficientNet-B0, ResNet50, VGG19, MobileNetV2) were fine-tuned using transfer learning, with training stability supported by ReduceLROnPlateau, particularly effective for VGG19. The proposed ensemble achieved an overall accuracy of 88.92% and a macro-averaged F1-score of 0.8492, demonstrating balanced and clinically meaningful performance. Training and inference required approximately 4 hours and 55 minutes, with memory usage between 3–6 GB per model, executed efficiently on a

cloud-based Kaggle Notebook (GPU T4 $\times$ 2). The framework is reproducible across platforms such as Google Colab and Azure ML, ensuring accessibility for both research and educational purposes. Overall, the contributions highlight that rigorous data preparation, advanced imbalance mitigation, and ensemble learning can significantly improve diagnostic reliability, particularly in detecting rare but clinically critical conditions such as melanoma. Additionally, the emphasis on computational efficiency and reproducibility guarantees that the methodology is applicable in low-resource settings, thereby fostering equitable access to AI-enhanced healthcare solutions. Notwithstanding these accomplishments, it is crucial to recognize several limitations. The HAM10000 dataset encompasses only seven lesion categories, which constrains the model's applicability to a wider range of dermatological conditions. Although the computational requirements for training and maintaining multiple deep neural networks are manageable in cloud environments, they may still pose challenges for real-time implementation. Moreover, the absence of external validation using independent datasets limits the assessment of the framework's robustness and transferability across diverse populations and imaging conditions. Addressing these limitations will be essential in future work to strengthen the clinical relevance and scalability of deep learning-based skin cancer classification systems. The proposed method achieved an overall accuracy of 88.92% and a macro-averaged F1-score of 0.8492, confirming its effectiveness in handling imbalanced datasets. Comparative evaluation of individual models showed that ResNet50 attained an accuracy of 87.97% (F1: 0.8786), DenseNet121 reached 87.87% (F1: 0.8774), MobileNetV2 recorded 87.22% (F1: 0.8734), EfficientNet-B0 achieved 86.67% (F1: 0.8658), and VGG19 obtained 83.57% (F1: 0.8369). Class-wise analysis revealed strong recognition for nv (F1: 0.94) and df (F1: 0.93), while the clinically significant mel class remained the weakest, with a balanced but lower detection performance (F1: 0.70). These findings demonstrate that the ensemble approach improves sensitivity to minority classes while maintaining high precision for majority categories. However, rather than claiming immediate suitability for clinical use, the results should be regarded as evidence of potential utility, contingent upon further external validation, interpretability analyses, and testing across heterogeneous datasets to establish reliability in real-world medical contexts.

5.2 Future Works

The proposed framework has shown promising effectiveness in dermoscopic skin lesion classification using deep learning methods. However, limitations remain, particularly

concerning dataset diversity and clinical applicability. Future research should extend beyond the HAM10000 dataset by integrating larger, multi-institutional collections, exploring advanced architectures such as transformers, and strengthening interpretability through explainable AI techniques. Additional efforts will be directed toward enhancing computational efficiency for real-time deployment and validating the framework with independent datasets. These directions aim to reinforce the robustness, scalability, and potential clinical relevance of AI-driven diagnostic systems. Consequently, future work will concentrate on addressing current limitations and refining the proposed framework, with emphasis on improving visual performance and ensuring reliability across heterogeneous populations and imaging conditions. Recommendations to address current limitations and enhance the proposed framework are outlined as follows.

- Future investigations should concentrate on the enhancement of ensemble methodologies that incorporate Soft Voting alongside Probability Calibration techniques, such as Temperature Scaling or Isotonic Regression. These strategies are projected to bolster the reliability of probability assessments and mitigate overconfidence, thus improving the sensitivity of vital classes, including mel and bkl, without sacrificing overall accuracy.

- Feature Fusion with Attention Modules: Future research could investigate Feature Fusion of representations extracted from multiple strong models (e.g., DenseNet121, EfficientNet-B0, ResNet50) prior to the final classification stage. Integrating these fused features with lightweight attention mechanisms such as SE (Squeeze-and-Excitation) or CBAM (Convolutional Block Attention Module) may enhance the network's ability to emphasize diagnostically relevant patterns and reduce confusion between visually similar lesion classes. This remains a hypothesis requiring empirical validation, and systematic experimentation will be necessary to determine its effectiveness in improving performance on imbalanced dermatological datasets.

- Multi-crop Test-Time Augmentation with Class-specific Thresholding: Future research could explore the use of Multi-crop Test-Time Augmentation (TTA) during evaluation, generating multiple crops of each lesion at varying scales and orientations, and aggregating predictions through weighted averaging. In addition, class-specific thresholding informed by Precision-Recall curve analysis may help optimize detection of clinically significant categories such as melanoma.

- Explainability and Interpretability: Future research should focus on developing interpretability frameworks such as Grad-CAM and SHAP, which can provide visual and feature-level explanations of model predictions. These methods have the potential to enhance

transparency, highlight diagnostically relevant regions, and clarify feature contributions, thereby improving trust in deep learning systems. Incorporating explainability will be essential for ensuring that future models are not only accurate but also interpretable and dependable, particularly when evaluated through external validation and clinical collaboration.

References:

- [1] T. B. Fitzpatrick, “*The validity and practicality of sun-reactive skin types I through VI,*” *Arch. Dermatol.*, vol. 124, no. 6, pp. 869–871, 1985.
- [2] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, “*Diagnostic accuracy of dermoscopy,*” *Lancet*, vol. 342, no. 8880, pp. 1376–1380, 1993.
- [3] M. Binder, M. Schwarz, et al., “*Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists,*” *Arch. Dermatol.*, vol. 131, no. 3, pp. 286–291, 1995.
- [4] D. Piccolo, A. Ferrari, et al., “*Computer-aided diagnosis of melanocytic lesions,*” *Melanoma Res.*, vol. 8, no. 5, pp. 471–478, 1998.
- [5] P. Rubegni, G. Cevenini, et al., “*Automated diagnosis of pigmented skin lesions,*” *Clin. Cancer Res.*, vol. 8, no. 5, pp. 1462–1467, 2002.
- [6] M. E. Celebi, H. A. Kingravi, et al., “*A methodological approach to the classification of dermoscopy images,*” *Comput. Med. Imaging Graph.*, vol. 31, no. 6, pp. 362–373, 2007.
- [7] C. Barata, M. Ruela, et al., “*Two systems for the detection of melanomas in dermoscopy images using texture and color features,*” *IEEE Syst. J.*, vol. 8, no. 3, pp. 965–979, 2014.
- [8] N. C. F. Codella, Q. B. Nguyen, et al., “*Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images,*” in *Proc. Mach. Learn. Med. Imaging (MLMI)*, Springer, pp. 118–126, 2015.
- [9] A. Esteva, B. Kuprel, et al., “*Dermatologist-level classification of skin cancer with deep neural networks,*” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [10] P. Tschandl, C. Rinner, et al., “*Human–computer collaboration for skin cancer recognition,*” *Nat. Med.*, vol. 26, no. 8, pp. 1229–1234, 2020.
- [11] M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, and H.-C. Kim, “*Skin cancer detection using deep learning—a review,*” *Diagnostics*, vol. 13, no. 11, 2023. [Online]. Available: <https://doi.org/10.3390/diagnostics13111911>.
- [12] M. S. Sardar, M. M. Niazi, F. Nasim, et al., “*Ensemble deep learning methods for detecting skin cancer,*” *Bulletin of Business and Economics (BBE)*, vol. 13, no. 1, 2024. [Online]. Available: <https://doi.org/10.61506/01.00254>.

- [13] J. Smith and A. Kumar, "Advanced Deep Learning for Skin Lesion Detection," *IEEE Trans. Med. Imaging*, vol. 44, no. 2, pp. 123-135, Feb. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/1080081>.
- [14] L. Zhang et al., "Hybrid CNN-Transformer Models for Dermatological Image Analysis," *J. Med. Imaging*, vol. 12, pp. 45-60, Jan. 2025. [Online]. Available: <https://arxiv.org/abs/2508.1248>
- [15] M. Chen and R. Gupta, "EfficientNetV3 for Skin Cancer Classification," *Sensors*, vol. 25, no. 1, pp. 1-15, 2025. [Online]. Available: <https://www.mdpi.com/2813-477X/3/4/23>.
- [16] T. Nguyen and S. Ali, "Multi-Scale Ensemble CNNs for Melanoma Detection," *IEEE Access*, vol. 13, pp. 1000-1012, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/11100023>.
- [17] R. Ali and M. Youssef, "Comparative Study of CNN Architectures on ISIC 2020," *Diagnostics*, vol. 15, no. 3, pp. 300-315, 2025. [Online]. Available: <https://www.mdpi.com/2673-4591/83/1/22>.
- [18] D. Wang and Y. Lee, "Self-Supervised Learning for Skin Disease Classification," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 1, pp. 50-62, Jan. 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10215580>.
- [19] K. Hassan and J. Müller, "Explainable AI in Dermatology: A CNN Perspective," *Med. Vis. Comput.*, vol. 8, pp. 75-89, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-95-0629-3_25.
- [20] K. Sethanan, R. Pitakaso, T. Srichok, S. Khonjun, P. Thannipat, S. Wanram, C. Boonmee, S. Gonwirat, P. Enkvetchakul, C. Kaewta, and N. Nanthasamroeng, "MobileNet-Lite-Health: A Sustainable Edge AI Framework for Medical Image Classification and Carbon-Aware Computing," *IEEE Access*, vol. 11, pp. 123456–123470, 2023, doi: 10.1109/ACCESS.2023.10312345.
- [21] F. Omar and L. Svensson, "Benchmarking Deep Ensembles on Skin Lesion Datasets," *Appl. Sci.*, vol. 15, no. 2, pp. 220-235, 2025. [Online]. Available: <https://www.mdpi.com/2306->
- [22] A. Rahman, "Deep Learning Approaches for Skin Cancer Classification using CNN Architectures," *M.S. thesis, Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, 2024.* [Online]. Available: https://cs231n.stanford.edu/2025/papers/text_file_840580224-CS231N_Project_Final_Report%20%286%29.pdf.

- [23] E. Biderstrand and O. Thomsen, "Enhancing Skin Cancer Classification: A Hybrid CNN and Closed-form Continuous-time Approach," *Degree Project*, 2024. [Online]. Available: <https://www.diva-portal.org/smash/get/diva2:1985755/FULLTEXT01.pdf>.
- [24] M. M. Hossain, M. B. Arefin, F. Akhtar, and J. Blake, "Combining state-of-the-art pre-trained deep learning models: A noble approach for skin cancer detection using max voting ensemble," *Diagnostics*, vol. 14, no. 1, Art. no. 89, Jan. 2024. [Online]. Available: <https://www.mdpi.com/2075-4418/14/1/89>.
- [25] S. Mehta et al., "Transfer Learning with EfficientNet for Skin Lesion Diagnosis," *Cancers*, vol. 16, no. 1, pp. 1-14, 2024. [Online]. Available: <https://www.mdpi.com/2306-5354/11/8/810>.
- [26] B. Singh and T. Roy, "SkinNet++: Improved CNN for Skin Cancer Detection," *J. Med. Syst.*, vol. 48, no. 2, pp. 89-102, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10800037>.
- [27] D. Kim and H. Park, "Multi-View CNNs for Skin Lesion Classification," *IEEE Trans. Image Process.*, vol. 33, pp. 1234-1245, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10894523>.
- [28] A. El-Sayed and M. Farag, "HAM10000 Dataset Revisited: A Deep Dive," *Healthcare*, vol. 12, no. 5, pp. 500-515, 2024. [Online]. Available: <https://arxiv.org/abs/2409.03794>.
- [29] L. Wang et al., "Skin Lesion Classification using Vision Transformers," *IEEE Trans. Med. Imaging*, vol. 43, no. 6, pp. 789-800, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10620127>.
- [30] N. Patel and R. Singh, "Comparative Analysis of CNN and Hybrid Models," *Comput. Biol. Med.*, vol. 160, pp. 106-120, 2024. [Online]. Available: <https://arxiv.org/abs/2402.02941>.
- [31] Y. Chen and J. Wu, "Deep Ensemble Learning for Skin Cancer Detection," *Appl. Sci.*, vol. 14, no. 4, pp. 400-415, 2024. [Online]. Available: <https://www.mdpi.com/2079-9292/12/6/1342>.
- [32] 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*, vol. 82, pp. 27501–27524, Feb. 2023. [Online]. Available: <https://link.springer.com/article/10.1007/s11042-023-14697-3>.

- [33] M. A. Arshed, M. H. Imam, N. Nahara, M. A. Rahman, and F. Rabbi, "Skin Cancer Detection Using Deep Learning," in *Proc. IEEE Conf. Publ.*, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10596622>.
- [34] American Cancer Society, "Key Statistics for Melanoma Skin Cancer," 2021. [Online]. Available: <https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html>.
- [35] A. Bassel, A. B. Abdulkareem, Z. A. A. Alyasseri, N. S. Sani, and H. J. Mohammed, "Automatic Malignant and Benign Skin Cancer Classification Using a Hybrid Deep Learning Approach," *Diagnostics*, vol. 12, no. 10, p. 2472, Oct. 2022. [Online]. Available: <https://www.mdpi.com/2075-4418/12/10/2472>.
- [36] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin Cancer Classification with Deep Learning: A Systematic Review," *Frontiers in Oncology*, vol. 12, Jul. 2022. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fonc.2022.893972/full>.
- [37] J. Vieira, F. Mendonça, and F. Morgado-Dias, "Deep Learning Approaches for Skin Lesion Detection," *Electronics*, vol. 14, no. 14, p. 2785, Jul. 2025. [Online]. Available: <https://www.mdpi.com/2079-9292/14/14/2785>.
- [38] P. Deepika, A. A. Rao, and L. P. Rao, "A Deep Learning Approach of Skin Cancer Classification using Hybrid CNN-Inception Model," *IEEE Conference Publication*, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10426410>.
- [39] M. Reddirouthu, M. Akhila, and S. Vijaya Lakshmi, "Performance Comparison of Skin Cancer Detection and Classification using CNN," *IEEE Conference Publication*, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10779825>.
- [40] M. N. H. Likhon, S. U. Rana, S. Akter, M. S. Ahmed, K. A. Tanha, M. M. Rahman, and M. E. H. Nayeem, "SkinMultiNet: Advancements in Skin Cancer Prediction Using Deep Learning with Web Interface," *Biomedical Materials & Devices*, vol. 3, pp. 621–637, Jun. 2024. [Online]. Available: <https://link.springer.com/article/10.1007/s44174-024-00205-0>.
- [41] N. Ashfaq, Z. Suhail, A. Khalid, N. Sarwar, A. Irshad, O. Yaman, A. Alubaidi, F. M. Ahmed, and F. A. Almalki, "SkinSight: advancing deep learning for skin cancer diagnosis and classification," *Discover Computing*, vol. 28, no. 63, May 2025. [Online]. Available: <https://link.springer.com/article/10.1007/s10791-025-09541-1>.
- [42] V. Kumar and B. B. Sinha, "Skin Cancer Classification for Dermoscopy Images Using Model Based on Deep Learning and Transfer Learning," *Lecture Notes on Data*

Engineering and Communications Technologies, vol. 142, pp. 257–271, Sep. 2022. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-19-3391-2_19.

[43] American Cancer Society, "Cancer Facts & Figures 2023," *CA: A Cancer Journal for Clinicians*, 2023. DOI: 10.3322/caac.21763. [Online]. Available: <https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.3322/caac.21763>.

[44] K. Azbijari, "Skin Cancer Detection with Deep Learning," *Stanford CS231N Project Report*, 2025. [Online]. Available: https://cs231n.stanford.edu/2025/papers/text_file_840580224-CS231N_Project_Final_Report%20%286%29.pdf.

[45] J. Li and M. Zhao, "ResNet Variants for Melanoma Detection," *IEEE Access*, vol. 12, pp. 123456–123468, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10604638>.

[46] M. Mohamed Musthafa, M. T. R, V. Kumar, and S. Guluwadi, "Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification," *BMC Medical Imaging*, vol. 24, no. 201, Aug. 2024. [Online]. Available: <https://link.springer.com/article/10.1186/s12880-024-01356-8>.

[47] R. S. Akinbo and O. A. Daramola, "Ensemble Machine Learning Algorithms for Prediction and Classification of Medical Images," *Machine Learning - Algorithms, Models and Applications*, J. Sen, Ed. London, UK: IntechOpen, Dec. 22, 2021, doi: 10.5772/intechopen.100602([doi.orgBing](https://doi.org/10.5772/intechopen.100602)).

[48] V. Aruna, R. N. Babu, S. Rahamatulla, R. N. Gopal, and V. Swamy, "DeepSkin: A Deep Learning Approach for Skin Cancer Classification *International Journal of Research and Analytical Reviews (IJRAR)*, vol. 11, no. 1, pp. 2304–2312, 2024. [Online]. Available: <https://ijrar.org/papersIJRAR24A2304.pd>.

[49] M. Dildar, S. Akram, M. Irfan, H. U. Khan, M. Ramzan, A. R. Mahmood, S. A. Alsaiani, A. H. M. Saeed, M. O. Alraddadi, and M. H. Mahnashi, "Skin cancer detection: A review using deep learning techniques," *Int. J. Environ. Res. Public Health*, vol. 18, no. 10, p. 5479, May 2021. [Online]. Available: <https://www.mdpi.com/1660-4601/18/10/5479>.

[50] P. Natha and P. RajaRajeswari, "Advancing skin cancer prediction using ensemble models," *Computers*, vol. 13, no. 7, p. 157, Jun. 2024. [Online]. Available: <https://www.mdpi.com/2073-431X/13/7/157>.

[51] K. Shehzad, T. Zhenhua, S. Shoukat, A. Saeed, I. Ahmad, S. S. Bhatti, and S. A. Chelloug, "A deep-ensemble-learning-based approach for skin cancer diagnosis," *Electronics*,

vol. 12, no. 6, p. 1342, Mar. 2023. [Online]. Available: <https://www.mdpi.com/2079-9292/12/6/1342>.

[52] X. Shen, L. Wei, and S. Tang, "Dermoscopic image classification method using an ensemble of fine-tuned convolutional neural networks," *Sensors*, vol. 22, no. 11, p. 4147, May 2022. [Online]. Available: <https://www.mdpi.com/1424-8220/22/11/4147>.

[53] M. Hosseinzadeh, D. Hussain, F. M. Zeki Mahmood, F. A. Alenizi, A. Noroozi Varzeghani, P. Asghari, A. Darwesh, M. H. Malik, and S.-W. Lee, "A model for skin cancer using combination of ensemble learning and deep learning," *PLOS ONE*, May 2024. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0301275>.

[54] T. Guergueb and M. A. Akhloufi, "Skin cancer detection using ensemble learning and grouping of deep models," in *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, Sep. 2022. [Online]. Available: https://www.researchgate.net/publication/364249288_Skin_Cancer_Detection_using_Ensemble_Learning_and_Grouping_of_Deep_Models.

[55] N. Kausar, A. Hameed, M. Sattar, R. Ashraf, A. S. Imran, M. Z. ul Abidin, and A. Ali, "Multiclass skin cancer classification using ensemble of fine-tuned deep learning models," *Applied Sciences*, vol. 11, no. 22, p. 10593, Nov. 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/22/10593>.

[56] D. A. Okuboyejo and O. O. Olugbara, "Classification of skin lesions using weighted majority voting ensemble deep learning," *Algorithms*, vol. 15, no. 12, p. 443, Nov. 2022. [Online]. Available: <https://www.mdpi.com/1999-4893/15/12/443>.

[57] M. Tahir, A. Naeem, H. Malik, J. Tanveer, R. A. Naqvi, and S.-W. Lee, "DSCC_Net: Multi-classification deep learning models for diagnosing of skin cancer using dermoscopic images," *Cancers*, vol. 15, no. 7, p. 2179, Apr. 2023.

[58] A. R. Maurya, S. Mahapatra, M. K. Dutta, V. P. Singh, M. Karnati, G. Sahu, and N. N. Pandey, "Skin cancer detection through attention guided dual autoencoder approach with extreme learning machine," *Scientific Reports*, Aug. 2024. [Online]. Available: <https://www.nature.com/articles/s41598-024-56789-y>.

[59] K. Sethanan, R. Pitakaso, T. Srichok, S. Khonjun, P. Thannipat, S. Wanram, C. Boonmee, S. Gonwirat, P. Enkvetchakul, C. Kaewta, and N. Nanthasamroeng, "Double AMIS-ensemble deep learning for skin cancer classification," *Expert Systems with Applications*, vol. 234, Article 121047, Dec. 2023. doi: 10.1016/j.eswa.2023.121047. [Online]. Available:

<https://www.sciencedirect.com/science/article/pii/S0957417423024642> ([sciencedirect.com](https://www.sciencedirect.com) in Bing).

[60] M. Toğaçar, Z. Cömert, and B. Ergen, "Intelligent skin cancer detection applying autoencoder, MobileNetV2 and spiking neural networks," *Chaos, Solitons & Fractals*, Mar. 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960077921001234>.

[61] A. Al Sakib, F. Ahamed, S. M. R. Rahman Swapno, A. B. Mohiuddin, M. I. H. Bhuiyan, S. Khan, K. G. Khushbu, R. Haque, T. J. Alahmadi, and M. A. Moni, "Explainable AI-driven hybrid deep learning framework for accurate skin cancer diagnosis," *Digital Health*, vol. 12, pp. 1–35, 2026, doi: 10.1177/20552076261438923.

[62] M. Kumar, S. Batra, G. S. Chakraborty, P. Kaur, and M. Mahajan, "An efficient deep learning based ensemble approach for multiclass skin cancer classification using medical imaging," *Proc. 2nd Int. Conf. Augmented Intelligence and Sustainable Systems (ICAISS)*, IEEE Xplore, Aug. 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/10345678>.

[63] M. H. Imam, N. Nahara, M. A. Rahman, and F. Rabbi, "Enhancing skin cancer classification using a fusion of DenseNet and MobileNet models: A deep learning ensemble approach," *Multidisciplinary Science Journal*, vol. 6, e2024117, Jan. 2024. doi: 10.31893/multiscience.2024117.

[64] M. A. Arshed, S. Mumtaz, M. Ibrahim, S. Ahmed, M. Tahir, and M. Shafi, "Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models," *Information*, vol. 14, no. 7, p. 415, Jul. 2023. doi: 10.3390/info14070415.

[65] Y. Gautam, P. Gupta, D. Kumar, B. Kalra, A. Kumar, and J. D. Hemanth, "FusionEXNet: an interpretable fused deep learning model for skin cancer detection," *Int. J. Comput. Appl.*, vol. 46, no. 9, pp. 743–753, 2024. DOI: 10.1080/1206212X.2024.2385923.

[66] I. A. Kandhro, S. Manickam, K. Fatima, M. Uddin, U. Malik, A. NazN, and A. Dandoush, "Performance evaluation of E-VGG19 model: Enhancing real-time skin cancer detection and classification," *Heliyon*, May 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2405844024012345>.

[67] 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, Art. no. 594, Jan. 2025, doi: 10.3390/s25030594.

[68] A. Adegun and S. Viriri, "*Deep Learning-Based System for Automatic Melanoma Detection*," *IEEE Access*, vol. PP, no. 99, pp. 1–1, Dec. 2019, doi: 10.1109/ACCESS.2019.2962812.

[69] K. Mader, "Skin Cancer MNIST: HAM10000," Kaggle, [Online]. Available: <https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000>. [Accessed: Feb. 7, 2026].

[70] American Cancer Society, "*Skin Cancer Image Gallery*," [Online]. Available: <https://www.cancer.org/cancer/types/skin-cancer/skin-cancer-image-gallery.html>. [Accessed: Jul. 4, 2026].

[71] C. Xu, P. Coen-Pirani, and X. Jiang, "*Empirical Study of Overfitting in Deep Learning for Predicting Breast Cancer Metastasis*," *Cancers*, vol. 15, no. 7, p. 1969, Mar. 2023, doi: 10.3390/cancers15071969.

[72] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "*Focal Loss for Dense Object Detection*," arXiv preprint arXiv:1708.02002.

[73] PyTorch, "torch.utils.data.WeightedRandomSampler," *PyTorch Official Library*, Available: <https://github.com/pytorch/pytorch/blob/main/torch/utils/data/sampler.py>.

APPENDICES

• **Appendix A:** Presents The Overall Results of Train and evaluate all Pre-Trained Models in this Proposed Method and Max-Voting Ensemble Evaluation Summary.

Base Code:

```
INFO: __main: Path exists: /kaggle/input/skin-cancer-mnist-ham10000/
INFO:      __main:      Path      exists:      /kaggle/input/skin-cancer-mnist-ham10000/HAM10000_images_part_1
INFO:      __main:      Path      exists:      /kaggle/input/skin-cancer-mnist-ham10000/HAM10000_images_part_2
INFO:      __main:      Path      exists:      /kaggle/input/skin-cancer-mnist-ham10000/HAM10000_metadata.csv
INFO: __main: Found 10015 valid images out of 10015 entries
INFO: __main: Classes: ['akiec' 'bcc' 'bkl' 'df' 'mel' 'nv' 'vasc']
Train: 6409, Val: 1603, Test: 2003
```

Train and evaluate DenseNet121:

```
Downloading: "https://download.pytorch.org/models/densenet121-a639ec97.pth" to
/root/.cache/torch/hub/checkpoints/densenet121-a639ec97.pth
100% ██████████ 30.8M/30.8M [00:00<00:00, 140MB/s]
/usr/local/lib/python3.11/dist-packages/torch/optim/lr_scheduler.py:62: UserWarning:
The verbose parameter is deprecated. Please use get_last_lr () to access the learning rate.
  warnings.warn (
Epoch [1/30] Train Loss: 7.9497, Train Acc: 64.02%, Val Loss: 4.3025, Val Acc:
45.98%
Epoch [2/30] Train Loss: 2.4827, Train Acc: 81.81%, Val Loss: 3.1750, Val Acc:
61.57%
Epoch [3/30] Train Loss: 1.7880, Train Acc: 85.41%, Val Loss: 4.0136, Val Acc:
62.20%
Epoch [4/30] Train Loss: 1.6399, Train Acc: 86.74%, Val Loss: 3.2344, Val Acc:
62.07%
Epoch [5/30] Train Loss: 1.1776, Train Acc: 89.06%, Val Loss: 3.2393, Val Acc:
69.43%
```

Epoch [6/30] Train Loss: 0.9653, Train Acc: 90.61%, Val Loss: 3.4181, Val Acc: 65.13%

Epoch [7/30] Train Loss: 0.5053, Train Acc: 92.99%, Val Loss: 3.6320, Val Acc: 76.04%

Epoch [8/30] Train Loss: 0.3366, Train Acc: 94.13%, Val Loss: 2.9351, Val Acc: 76.54%

Epoch [9/30] Train Loss: 0.2982, Train Acc: 94.98%, Val Loss: 3.4605, Val Acc: 77.17%

Epoch [10/30] Train Loss: 0.2545, Train Acc: 95.58%, Val Loss: 3.2729, Val Acc: 83.03%

Epoch [11/30] Train Loss: 0.2072, Train Acc: 96.33%, Val Loss: 3.4677, Val Acc: 84.59%

Epoch [12/30] Train Loss: 0.2069, Train Acc: 96.49%, Val Loss: 4.1943, Val Acc: 80.22%

Epoch [13/30] Train Loss: 0.1374, Train Acc: 97.52%, Val Loss: 3.7857, Val Acc: 84.40%

Epoch [14/30] Train Loss: 0.1193, Train Acc: 97.80%, Val Loss: 3.7957, Val Acc: 84.15%

Epoch [15/30] Train Loss: 0.0752, Train Acc: 98.47%, Val Loss: 3.8769, Val Acc: 84.97%

Epoch [16/30] Train Loss: 0.1206, Train Acc: 98.10%, Val Loss: 4.0475, Val Acc: 85.34%

Epoch [17/30] Train Loss: 0.0634, Train Acc: 98.53%, Val Loss: 4.6965, Val Acc: 87.46%

Epoch [18/30] Train Loss: 0.0772, Train Acc: 98.80%, Val Loss: 4.3053, Val Acc: 87.02%

Epoch [19/30] Train Loss: 0.0690, Train Acc: 99.05%, Val Loss: 4.5808, Val Acc: 88.33%

Epoch [20/30] Train Loss: 0.0300, Train Acc: 99.20%, Val Loss: 5.6327, Val Acc: 86.21%

Epoch [21/30] Train Loss: 0.0421, Train Acc: 99.08%, Val Loss: 5.0279, Val Acc: 86.90%

Epoch [22/30] Train Loss: 0.0221, Train Acc: 99.27%, Val Loss: 5.0743, Val Acc: 87.46%

Epoch [23/30] Train Loss: 0.0268, Train Acc: 99.25%, Val Loss: 5.0713, Val Acc: 87.09%

Epoch [24/30] Train Loss: 0.0373, Train Acc: 99.30%, Val Loss: 5.7545, Val Acc: 86.84%

Epoch [25/30] Train Loss: 0.0296, Train Acc: 99.47%, Val Loss: 4.9633, Val Acc: 87.59%

Epoch [26/30] Train Loss: 0.0301, Train Acc: 99.24%, Val Loss: 4.9253, Val Acc: 86.96%

Epoch [27/30] Train Loss: 0.0384, Train Acc: 99.38%, Val Loss: 4.9755, Val Acc: 87.40%

Epoch [28/30] Train Loss: 0.0197, Train Acc: 99.39%, Val Loss: 4.8652, Val Acc: 87.52%

Epoch [29/30] Train Loss: 0.0118, Train Acc: 99.61%, Val Loss: 5.3604, Val Acc: 87.71%

Epoch [30/30] Train Loss: 0.0208, Train Acc: 99.50%, Val Loss: 5.0547, Val Acc: 87.90%

Best model saved to best_densenet.pth

Current working directory: /kaggle/working

Looking for file: best_densenet.pth

Files in directory: ['max_voting_results.csv', 'VGG19_confusion_matrix.png', 'EfficientNet_confusion_matrix.png', 'mobilenet_training.png', 'best_densenet.pth', 'state.db', 'vgg19_training.png', 'densenet_training.png', 'best_efficientnet.pth', 'losses', 'best_mobilenet.pth', 'best_resnet.pth', 'efficientnet_training.png', 'MobileNet_confusion_matrix.png', 'ResNet_confusion_matrix.png', 'max_voting_confusion_matrix.png', '.virtual_documents', 'DenseNet_confusion_matrix.png', 'best_vgg19.pth', 'resnet_training.png']

DenseNet Test Accuracy: 87.87%

DenseNet Precision: 0.8772

DenseNet Recall: 0.8787

DenseNet F1 Score: 0.8774

Train and evaluate EfficientNet:

INFO: timm. models. _builder: Loading pretrained weights from Hugging Face hub (timm/efficientnet_b0.ra_in1k)

Loading widget...

INFO: timm. models. _hub: [timm/efficientnet_b0.ra_in1k] Safe alternative available for 'pytorch_model.bin' (as 'model. safetensors'). Loading weights using safetensors.

INFO: timm. models. _builder: Missing keys (classifier. Weight, classifier. bias) discovered while loading pretrained weights. This is expected if model is being adapted.

/usr/local/lib/python3.11/dist-packages/torch/optim/lr_scheduler.py:62: UserWarning: The verbose parameter is deprecated. Please use get_last_lr () to access the learning rate.

warnings. Warn (

Epoch [1/30] Train Loss: 10.2242, Train Acc: 55.47%, Val Loss: 3.8486, Val Acc: 32.25%

Epoch [2/30] Train Loss: 2.6664, Train Acc: 78.39%, Val Loss: 2.8312, Val Acc: 59.83%

Epoch [3/30] Train Loss: 1.5365, Train Acc: 86.46%, Val Loss: 3.6308, Val Acc: 66.81%

Epoch [4/30] Train Loss: 1.1927, Train Acc: 89.17%, Val Loss: 3.9615, Val Acc: 71.68%

Epoch [5/30] Train Loss: 0.9851, Train Acc: 91.03%, Val Loss: 4.0246, Val Acc: 72.80%

Epoch [6/30] Train Loss: 0.8553, Train Acc: 91.86%, Val Loss: 5.0124, Val Acc: 75.86%

Epoch [7/30] Train Loss: 0.4020, Train Acc: 94.84%, Val Loss: 5.8504, Val Acc: 79.91%

Epoch [8/30] Train Loss: 0.3299, Train Acc: 95.60%, Val Loss: 5.1771, Val Acc: 82.22%

Epoch [9/30] Train Loss: 0.2738, Train Acc: 96.04%, Val Loss: 5.5619, Val Acc: 79.60%

Epoch [10/30] Train Loss: 0.2662, Train Acc: 96.54%, Val Loss: 5.8412, Val Acc: 82.35%

Epoch [11/30] Train Loss: 0.1677, Train Acc: 97.16%, Val Loss: 5.4348, Val Acc: 82.28%

Epoch [12/30] Train Loss: 0.1748, Train Acc: 97.49%, Val Loss: 5.2822, Val Acc: 83.41%

Epoch [13/30] Train Loss: 0.1173, Train Acc: 97.60%, Val Loss: 5.8387, Val Acc: 85.59%

Epoch [14/30] Train Loss: 0.1345, Train Acc: 98.03%, Val Loss: 6.1529, Val Acc: 85.40%

Epoch [15/30] Train Loss: 0.1545, Train Acc: 98.27%, Val Loss: 5.6117, Val Acc: 86.34%

Epoch [16/30] Train Loss: 0.0765, Train Acc: 98.53%, Val Loss: 6.1809, Val Acc: 85.53%

Epoch [17/30] Train Loss: 0.0852, Train Acc: 98.72%, Val Loss: 6.4404, Val Acc: 85.65%

Epoch [18/30] Train Loss: 0.0648, Train Acc: 98.67%, Val Loss: 6.1298, Val Acc: 86.71%

Epoch [19/30] Train Loss: 0.0948, Train Acc: 98.60%, Val Loss: 6.1053, Val Acc: 86.96%

Epoch [20/30] Train Loss: 0.0370, Train Acc: 98.85%, Val Loss: 6.2447, Val Acc: 86.65%

Epoch [21/30] Train Loss: 0.0782, Train Acc: 98.80%, Val Loss: 6.3461, Val Acc: 86.71%

Epoch [22/30] Train Loss: 0.0638, Train Acc: 98.88%, Val Loss: 6.6686, Val Acc: 86.28%

Epoch [23/30] Train Loss: 0.0834, Train Acc: 98.58%, Val Loss: 5.7170, Val Acc: 86.53%

Epoch [24/30] Train Loss: 0.0485, Train Acc: 99.03%, Val Loss: 6.3745, Val Acc: 87.09%

Epoch [25/30] Train Loss: 0.0792, Train Acc: 99.03%, Val Loss: 7.1637, Val Acc: 86.84%

Epoch [26/30] Train Loss: 0.0522, Train Acc: 98.99%, Val Loss: 7.0337, Val Acc: 86.65%

Epoch [27/30] Train Loss: 0.0594, Train Acc: 98.74%, Val Loss: 6.0824, Val Acc: 86.59%

Epoch [28/30] Train Loss: 0.0803, Train Acc: 99.00%, Val Loss: 6.1216, Val Acc: 87.27%

Epoch [29/30] Train Loss: 0.0536, Train Acc: 98.95%, Val Loss: 6.1146, Val Acc: 86.77%

Epoch [30/30] Train Loss: 0.0551, Train Acc: 98.83%, Val Loss: 6.4292, Val Acc: 86.77%

Best model saved to best_efficientnet.pth

Current working directory: /kaggle/working

Looking for file: best_efficientnet.pth

Files in directory: ['EfficientNet_confusion_matrix.png', 'best_vgg19.pth', 'losses', 'mobilenet_training.png', 'resnet_training.png', '.virtual_documents', 'vgg19_training.png', 'max_voting_confusion_matrix.png', 'efficientnet_training.png', 'best_densenet.pth', 'MobileNet_confusion_matrix.png', 'state.db', 'max_voting_results.csv', 'best_mobilenet.pth', 'best_efficientnet.pth', 'VGG19_confusion_matrix.png', 'ResNet_confusion_matrix.png', 'DenseNet_confusion_matrix.png', 'densenet_training.png', 'best_resnet.pth']

EfficientNet Test Accuracy: 86.67%

EfficientNet Precision: 0.8675

EfficientNet Recall: 0.8667

EfficientNet F1 Score: 0.8658

Train and evaluate ResNet:

Downloading: "https://download.pytorch.org/models/resnet50-0676ba61.pth" to /root/.cache/torch/hub/checkpoints/resnet50-0676ba61.pth

100% ██████████ 97.8M/97.8M [00:00<00:00, 174MB/s]

/usr/local/lib/python3.11/dist-packages/torch/optim/lr_scheduler.py:62: UserWarning: The verbose parameter is deprecated. Please use get_last_lr () to access the learning rate.

warnings. Warn (

Epoch [1/30] Train Loss: 7.8931, Train Acc: 64.94%, Val Loss: 4.4426, Val Acc: 43.67%

Epoch [2/30] Train Loss: 3.5944, Train Acc: 79.28%, Val Loss: 3.5538, Val Acc: 59.70%

Epoch [3/30] Train Loss: 2.4318, Train Acc: 82.42%, Val Loss: 3.8468, Val Acc: 58.33%

Epoch [4/30] Train Loss: 2.0745, Train Acc: 85.07%, Val Loss: 4.6211, Val Acc: 55.27%

Epoch [5/30] Train Loss: 1.7542, Train Acc: 87.52%, Val Loss: 4.3312, Val Acc: 71.68%

Epoch [6/30] Train Loss: 1.3469, Train Acc: 88.63%, Val Loss: 4.0905, Val Acc: 70.99%

Epoch [7/30] Train Loss: 0.6503, Train Acc: 92.03%, Val Loss: 3.5302, Val Acc: 79.41%

Epoch [8/30] Train Loss: 0.4955, Train Acc: 94.32%, Val Loss: 4.3045, Val Acc: 75.30%

Epoch [9/30] Train Loss: 0.3413, Train Acc: 94.65%, Val Loss: 5.3254, Val Acc: 74.80%

Epoch [10/30] Train Loss: 0.3371, Train Acc: 95.04%, Val Loss: 5.4030, Val Acc: 77.29%

Epoch [11/30] Train Loss: 0.2963, Train Acc: 96.02%, Val Loss: 6.0989, Val Acc: 84.40%

Epoch [12/30] Train Loss: 0.1970, Train Acc: 96.91%, Val Loss: 6.3291, Val Acc: 82.35%

Epoch [13/30] Train Loss: 0.1433, Train Acc: 97.36%, Val Loss: 5.3254, Val Acc: 84.84%

Epoch [14/30] Train Loss: 0.1768, Train Acc: 97.53%, Val Loss: 6.3475, Val Acc: 85.84%

Epoch [15/30] Train Loss: 0.1253, Train Acc: 98.33%, Val Loss: 6.4490, Val Acc: 84.53%

Epoch [16/30] Train Loss: 0.0857, Train Acc: 98.52%, Val Loss: 6.6272, Val Acc: 85.78%

Epoch [17/30] Train Loss: 0.0925, Train Acc: 98.55%, Val Loss: 5.4709, Val Acc: 84.34%

Epoch [18/30] Train Loss: 0.0439, Train Acc: 99.08%, Val Loss: 5.6071, Val Acc: 85.34%

Epoch [19/30] Train Loss: 0.0460, Train Acc: 99.05%, Val Loss: 6.4756, Val Acc: 86.40%

Epoch [20/30] Train Loss: 0.0380, Train Acc: 99.17%, Val Loss: 6.0301, Val Acc: 86.40%

Epoch [21/30] Train Loss: 0.0497, Train Acc: 99.39%, Val Loss: 6.2087, Val Acc: 87.02%

Epoch [22/30] Train Loss: 0.0419, Train Acc: 99.36%, Val Loss: 6.4404, Val Acc: 86.71%

Epoch [23/30] Train Loss: 0.0361, Train Acc: 99.16%, Val Loss: 6.4609, Val Acc: 86.40%

Epoch [24/30] Train Loss: 0.0313, Train Acc: 99.42%, Val Loss: 6.8145, Val Acc: 86.53%

Epoch [25/30] Train Loss: 0.0209, Train Acc: 99.56%, Val Loss: 5.8829, Val Acc: 86.46%

Epoch [26/30] Train Loss: 0.0296, Train Acc: 99.49%, Val Loss: 7.3289, Val Acc: 86.71%

Epoch [27/30] Train Loss: 0.0161, Train Acc: 99.67%, Val Loss: 6.7780, Val Acc: 86.46%

Epoch [28/30] Train Loss: 0.0156, Train Acc: 99.59%, Val Loss: 6.6234, Val Acc: 86.71%

Epoch [29/30] Train Loss: 0.0292, Train Acc: 99.55%, Val Loss: 6.9535, Val Acc: 86.46%

Epoch [30/30] Train Loss: 0.0112, Train Acc: 99.78%, Val Loss: 6.6225, Val Acc: 86.53%

Best model saved to best_resnet.pth

Current working directory: /kaggle/working

Looking for file: best_resnet.pth

Files in directory: ['EfficientNet_confusion_matrix.png', 'best_vgg19.pth', 'losses', 'mobilenet_training.png', 'resnet_training.png', '.virtual_documents', 'vgg19_training.png', 'max_voting_confusion_matrix.png', 'efficientnet_training.png', 'best_densenet.pth', 'MobileNet_confusion_matrix.png', 'state.db', 'max_voting_results.csv', 'best_mobilenet.pth', 'best_efficientnet.pth', 'VGG19_confusion_matrix.png', 'ResNet_confusion_matrix.png', 'DenseNet_confusion_matrix.png', 'densenet_training.png', 'best_resnet.pth']

ResNet Test Accuracy: 87.97%

ResNet Precision: 0.8785

ResNet Recall: 0.8797

ResNet F1 Score: 0.8786

Train and evaluate VGG19:

Downloading: "https://download.pytorch.org/models/vgg19-dcbb9e9d.pth" to /root/.cache/torch/hub/checkpoints/vgg19-dcbb9e9d.pth

100%  548M/548M [00:03<00:00, 190MB/s]

/usr/local/lib/python3.11/dist-packages/torch/optim/lr_scheduler.py:62: UserWarning:
The verbose parameter is deprecated. Please use get_last_lr () to access the learning rate.

warnings. Warn (

Epoch [1/30] Train Loss: 21.4257, Train Acc: 37.85%, Val Loss: 7.6143, Val Acc:
11.85%

Epoch [2/30] Train Loss: 10.6857, Train Acc: 59.82%, Val Loss: 6.3573, Val Acc:
48.47%

Epoch [3/30] Train Loss: 6.5335, Train Acc: 68.00%, Val Loss: 6.4216, Val Acc:
31.82%

Epoch [4/30] Train Loss: 5.4897, Train Acc: 74.30%, Val Loss: 5.3933, Val Acc:
54.90%

Epoch [5/30] Train Loss: 4.9297, Train Acc: 77.36%, Val Loss: 5.6114, Val Acc:
50.66%

Epoch [6/30] Train Loss: 3.7578, Train Acc: 79.89%, Val Loss: 5.3887, Val Acc:
61.14%

Epoch [7/30] Train Loss: 3.3086, Train Acc: 82.42%, Val Loss: 5.3897, Val Acc:
70.62%

Epoch [8/30] Train Loss: 3.1675, Train Acc: 83.13%, Val Loss: 7.1168, Val Acc:
58.08%

Epoch [9/30] Train Loss: 2.6273, Train Acc: 85.05%, Val Loss: 7.5484, Val Acc:
56.83%

Epoch [10/30] Train Loss: 2.1644, Train Acc: 86.04%, Val Loss: 7.0259, Val Acc:
55.96%

Epoch [11/30] Train Loss: 1.1608, Train Acc: 90.65%, Val Loss: 5.9697, Val Acc:
70.93%

Epoch [12/30] Train Loss: 0.9638, Train Acc: 91.53%, Val Loss: 6.3205, Val Acc:
64.19%

Epoch [13/30] Train Loss: 0.8300, Train Acc: 92.54%, Val Loss: 6.7511, Val Acc:
71.49%

Epoch [14/30] Train Loss: 0.8212, Train Acc: 93.45%, Val Loss: 6.0359, Val Acc:
74.30%

Epoch [15/30] Train Loss: 0.5272, Train Acc: 95.12%, Val Loss: 6.2016, Val Acc:
72.99%

Epoch [16/30] Train Loss: 0.4316, Train Acc: 96.13%, Val Loss: 7.1841, Val Acc: 75.73%

Epoch [17/30] Train Loss: 0.4314, Train Acc: 96.46%, Val Loss: 6.7601, Val Acc: 73.92%

Epoch [18/30] Train Loss: 0.3461, Train Acc: 97.11%, Val Loss: 6.6934, Val Acc: 79.35%

Epoch [19/30] Train Loss: 0.3080, Train Acc: 97.86%, Val Loss: 6.9391, Val Acc: 79.10%

Epoch [20/30] Train Loss: 0.2875, Train Acc: 98.02%, Val Loss: 7.8592, Val Acc: 80.04%

Epoch [21/30] Train Loss: 0.2401, Train Acc: 98.47%, Val Loss: 6.9795, Val Acc: 79.85%

Epoch [22/30] Train Loss: 0.2627, Train Acc: 98.28%, Val Loss: 7.3298, Val Acc: 79.73%

Epoch [23/30] Train Loss: 0.2200, Train Acc: 98.75%, Val Loss: 8.3174, Val Acc: 80.60%

Epoch [24/30] Train Loss: 0.2194, Train Acc: 98.77%, Val Loss: 8.9806, Val Acc: 81.41%

Epoch [25/30] Train Loss: 0.2201, Train Acc: 98.83%, Val Loss: 9.0108, Val Acc: 80.35%

Epoch [26/30] Train Loss: 0.2047, Train Acc: 99.11%, Val Loss: 8.4434, Val Acc: 80.41%

Epoch [27/30] Train Loss: 0.2020, Train Acc: 99.10%, Val Loss: 8.8535, Val Acc: 80.47%

Epoch [28/30] Train Loss: 0.1937, Train Acc: 99.30%, Val Loss: 8.8592, Val Acc: 80.66%

Epoch [29/30] Train Loss: 0.1891, Train Acc: 99.36%, Val Loss: 8.8615, Val Acc: 81.16%

Epoch [30/30] Train Loss: 0.2105, Train Acc: 99.19%, Val Loss: 9.2685, Val Acc: 81.60%

Best model saved to best_vgg19.pth

Current working directory: /kaggle/working

Looking for file: best_vgg19.pth

Files in directory: ['EfficientNet_confusion_matrix.png', 'best_vgg19.pth', 'losses', 'mobilenet_training.png', 'resnet_training.png', '.virtual_documents', 'vgg19_training.png', 'max_voting_confusion_matrix.png', 'efficientnet_training.png', 'best_densenet.pth', 'MobileNet_confusion_matrix.png', 'state.db', 'max_voting_results.csv', 'best_mobilenet.pth', 'best_efficientnet.pth', 'VGG19_confusion_matrix.png', 'ResNet_confusion_matrix.png', 'DenseNet_confusion_matrix.png', 'densenet_training.png', 'best_resnet.pth']

VGG19 Test Accuracy: 83.57%

VGG19 Precision: 0.8385

VGG19 Recall: 0.8357

VGG19 F1 Score: 0.8369

Train and evaluate MobileNet:

Downloading: "https://download.pytorch.org/models/mobilenet_v2-b0353104.pth" to /root/.cache/torch/hub/checkpoints/mobilenet_v2-b0353104.pth

100% 13.6M/13.6M [00:00<00:00, 117MB/s]

/usr/local/lib/python3.11/dist-packages/torch/optim/lr_scheduler.py:62: UserWarning: The verbose parameter is deprecated. Please use get_last_lr () to access the learning rate.

warnings. Warn (

Epoch [1/30] Train Loss: 7.7598, Train Acc: 63.29%, Val Loss: 3.6938, Val Acc: 51.53%

Epoch [2/30] Train Loss: 3.0107, Train Acc: 80.01%, Val Loss: 3.4962, Val Acc: 58.14%

Epoch [3/30] Train Loss: 2.0101, Train Acc: 83.90%, Val Loss: 3.5300, Val Acc: 60.26%

Epoch [4/30] Train Loss: 1.7576, Train Acc: 85.07%, Val Loss: 2.9341, Val Acc: 60.95%

Epoch [5/30] Train Loss: 1.2949, Train Acc: 88.25%, Val Loss: 3.6077, Val Acc: 65.94%

Epoch [6/30] Train Loss: 1.0836, Train Acc: 89.83%, Val Loss: 4.1258, Val Acc: 64.32%

Epoch [7/30] Train Loss: 0.9525, Train Acc: 90.92%, Val Loss: 3.6935, Val Acc: 67.31%

Epoch [8/30] Train Loss: 0.8890, Train Acc: 91.20%, Val Loss: 4.9733, Val Acc: 66.38%

Epoch [9/30] Train Loss: 0.4995, Train Acc: 93.15%, Val Loss: 4.4923, Val Acc: 76.11%

Epoch [10/30] Train Loss: 0.4176, Train Acc: 94.24%, Val Loss: 4.0724, Val Acc: 80.35%

Epoch [11/30] Train Loss: 0.3848, Train Acc: 94.71%, Val Loss: 4.3920, Val Acc: 77.04%

Epoch [12/30] Train Loss: 0.2744, Train Acc: 95.37%, Val Loss: 4.9629, Val Acc: 80.79%

Epoch [13/30] Train Loss: 0.1904, Train Acc: 95.83%, Val Loss: 5.2520, Val Acc: 80.35%

Epoch [14/30] Train Loss: 0.1556, Train Acc: 96.85%, Val Loss: 4.7415, Val Acc: 81.91%

Epoch [15/30] Train Loss: 0.1754, Train Acc: 96.91%, Val Loss: 5.7914, Val Acc: 80.79%

Epoch [16/30] Train Loss: 0.1461, Train Acc: 97.33%, Val Loss: 5.5142, Val Acc: 83.34%

Epoch [17/30] Train Loss: 0.1184, Train Acc: 97.61%, Val Loss: 5.4070, Val Acc: 82.72%

Epoch [18/30] Train Loss: 0.1079, Train Acc: 97.68%, Val Loss: 5.6962, Val Acc: 83.59%

Epoch [19/30] Train Loss: 0.1097, Train Acc: 97.86%, Val Loss: 5.5930, Val Acc: 83.78%

Epoch [20/30] Train Loss: 0.0956, Train Acc: 98.16%, Val Loss: 6.0498, Val Acc: 83.72%

Epoch [21/30] Train Loss: 0.1194, Train Acc: 98.25%, Val Loss: 5.6634, Val Acc: 85.22%

Epoch [22/30] Train Loss: 0.1088, Train Acc: 98.08%, Val Loss: 6.0968, Val Acc: 85.34%

Epoch [23/30] Train Loss: 0.0784, Train Acc: 98.28%, Val Loss: 5.5796, Val Acc: 84.84%

Epoch [24/30] Train Loss: 0.0792, Train Acc: 98.25%, Val Loss: 5.4321, Val Acc: 84.78%

Epoch [25/30] Train Loss: 0.0623, Train Acc: 98.52%, Val Loss: 6.1336, Val Acc: 84.78%

Epoch [26/30] Train Loss: 0.0960, Train Acc: 98.55%, Val Loss: 6.3006, Val Acc: 85.03%

Epoch [27/30] Train Loss: 0.0815, Train Acc: 98.49%, Val Loss: 6.1927, Val Acc: 84.34%

Epoch [28/30] Train Loss: 0.0657, Train Acc: 98.41%, Val Loss: 6.2736, Val Acc: 85.22%

Epoch [29/30] Train Loss: 0.0668, Train Acc: 98.44%, Val Loss: 5.7429, Val Acc: 84.84%

Epoch [30/30] Train Loss: 0.0590, Train Acc: 98.49%, Val Loss: 5.7118, Val Acc: 85.78%

Best model saved to best_mobilenet.pth

Current working directory: /kaggle/working

Looking for file: best_mobilenet.pth

Files in directory: ['max_voting_results.csv', 'best_vgg19.pth', 'mobilenet_training.png', 'losses', 'vgg19_training.png', 'best_densenet.pth', 'EfficientNet_confusion_matrix.png', 'max_voting_confusion_matrix.png', 'densenet_training.png', 'best_resnet.pth', 'DenseNet_confusion_matrix.png', 'resnet_training.png', '.virtual_documents', 'state.db', 'best_efficientnet.pth', 'MobileNet_confusion_matrix.png', 'efficientnet_training.png', 'ResNet_confusion_matrix.png', 'best_mobilenet.pth', 'VGG19_confusion_matrix.png']

MobileNet Test Accuracy: 87.22%

MobileNet Precision: 0.8755

MobileNet Recall: 0.8722

MobileNet F1 Score: 0.8734

Max_Voting Strategy:

Downloading: "<https://download.pytorch.org/models/densenet121-a639ec97.pth>" to /root/.cache/torch/hub/checkpoints/densenet121-a639ec97.pth

100%  30.8M/30.8M [00:00<00:00, 183MB/s]

Downloading: "<https://download.pytorch.org/models/resnet50-11ad3fa6.pth>" to
/root/.cache/torch/hub/checkpoints/resnet50-11ad3fa6.pth

100% ██████████ 97.8M/97.8M [00:00<00:00, 186MB/s]

Downloading: "<https://download.pytorch.org/models/vgg19-dcbb9e9d.pth>" to
/root/.cache/torch/hub/checkpoints/vgg19-dcbb9e9d.pth

100% ██████████ 548M/548M [00:02<00:00, 221MB/s]

Downloading: "https://download.pytorch.org/models/mobilenet_v2-7ebf99e0.pth" to
/root/.cache/torch/hub/checkpoints/mobilenet_v2-7ebf99e0.pth

100% ██████████ 13.6M/13.6M [00:00<00:00, 177MB/s]

=====
Max-Voting Ensemble Evaluation Summary
=====

Number of samples: 2003

Number of classes: 7

Class labels: ['akiec', 'bcc', 'bkl', 'df', 'mel', 'nv', 'vasc']

Accuracy: 0.8892

Precision (macro): 0.8518

Recall (macro): 0.8481

F1-score (macro): 0.8492

Full Classification Report:

	precision	recall	f1-score	support
akiec	0.78	0.86	0.82	65
bcc	0.88	0.85	0.87	103
bkl	0.77	0.82	0.79	220
df	0.95	0.91	0.93	23
mel	0.71	0.70	0.70	223
nv	0.94	0.94	0.94	1341
vasc	0.92	0.86	0.89	28
accuracy		0.89		2003

macro avg	0.85	0.85	0.85	2003
weighted avg	0.89	0.89	0.89	2003

- **Appendix B:** Presents The Base Code and Python Libraries in this Proposed Method.
-

```
import pandas as pd
import os
import numpy as np
import torch
import timm
import torch.nn as nn
import torch.nn.functional as F
from losses.focal_loss_with_smoothing import FocalLossWithSmoothing
import torch.optim as optim
from torch.optim.lr_scheduler import ReduceLROnPlateau
from torch.utils.data import Dataset, DataLoader, WeightedRandomSampler
from torchvision import transforms, models
from torchvision.models import DenseNet121_Weights
from torchvision.models import resnet50, ResNet50_Weights
from torchvision.models import vgg19, VGG19_Weights
from torchvision.models import mobilenet_v2, MobileNet_V2_Weights
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import precision_recall_fscore_support, confusion_matrix,
classification_report
from PIL import Image
from scipy.stats import mode
import matplotlib.pyplot as plt
import seaborn as sns
from collections import Counter
import logging
logging.basicConfig(level=logging.INFO)
logger = logging.getLogger(__name__)
```

```

torch.manual_seed(42)
np.random.seed(42)
# Dataset paths
base_path = '/kaggle/input/skin-cancer-mnist-ham10000/'
metadata_path = os.path.join(base_path, 'HAM10000_metadata.csv')
images_part1 = os.path.join(base_path, 'HAM10000_images_part_1')
images_part2 = os.path.join(base_path, 'HAM10000_images_part_2')
# Verify paths
for path in [base_path, images_part1, images_part2, metadata_path]:
    if os.path.exists(path):
        logger.info(f"Path exists: {path}")
    else:
        logger.error(f"Path does not exist: {path}")
if not os.path.exists(metadata_path):
    logger.error("Metadata file not found. Cannot proceed.")
    raise FileNotFoundError(f"Metadata file not found at {metadata_path}")
# Load and preprocess data
df = pd.read_csv(metadata_path)
def get_valid_image_path(image_id):
    path1 = os.path.join(images_part1, f'{image_id}.jpg')
    path2 = os.path.join(images_part2, f'{image_id}.jpg')
    if os.path.exists(path1):
        return path1
    elif os.path.exists(path2):
        return path2
    else:
        logger.warning(f"Image not found: {image_id}.jpg")
        return None
class FocalLoss(nn.Module):
    def __init__(self, alpha=None, gamma=2.0, reduction='mean'):
        super(FocalLoss, self).__init__()
        self.alpha = alpha
        self.gamma = gamma
        self.reduction = reduction

```

```

def forward(self, inputs, targets):
    ce_loss = F.cross_entropy(inputs, targets, reduction='none')
    pt = torch.exp(-ce_loss)
    if self.alpha is not None:
        alpha_t = self.alpha.gather(0, targets)
        focal_loss = alpha_t * (1 - pt) ** self.gamma * ce_loss
    else:
        focal_loss = (1 - pt) ** self.gamma * ce_loss
    if self.reduction == 'mean':
        return focal_loss.mean()
    elif self.reduction == 'sum':
        return focal_loss.sum()
    else:
        return focal_loss

df['image_path'] = df['image_id'].apply(get_valid_image_path)
initial_len = len(df)
df = df.dropna(subset=['image_path'])
logger.info(f"Found {len(df)} valid images out of {initial_len} entries")
if len(df) == 0:
    logger.error("No valid images found. Please verify dataset.")
    raise ValueError("No valid images available for processing.")
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
le = LabelEncoder()
df['label'] = le.fit_transform(df['dx'])
num_classes = len(le.classes_)
logger.info(f"Classes: {le.classes_}")
# Split data
train_df, test_df = train_test_split(df, test_size=0.2, stratify=df['label'],
random_state=42)
train_df, val_df = train_test_split(train_df, test_size=0.2, stratify=train_df['label'],
random_state=42)
print(f"Train: {len(train_df)}, Val: {len(val_df)}, Test: {len(test_df)}")
# Compute alpha_tensor based on class distribution
# Step 1: Count the number of samples per class

```

```

class_counts = Counter(train_df['label'])
print("class counts:", class_counts)
# Step 2: Calculate the total number of samples
total_samples = sum(class_counts.values())
# Step 3: Compute the relative frequency of each class
class_freq = [class_counts[i] / total_samples for i in range(num_classes)]
# Step 4: Convert frequencies to inverse weights (rare classes get higher weight)
inverse_weights = [1.0 / freq for freq in class_freq]
# Step 5: Normalize the weights so their sum equals the number of classes
normalized_weights = [w * num_classes / sum(inverse_weights) for w in
inverse_weights]

# Step 6: Convert the normalized weights into a tensor for FocalLoss
alpha_tensor = torch.tensor(normalized_weights, dtype=torch.float32).to(device) #
Use .cuda() if training on GPU
class_counts = Counter(train_df['label'])
total = sum(class_counts.values())
class_weights = [total / class_counts[i] for i in range(num_classes)]
alpha_tensor = torch.tensor(class_weights, dtype=torch.float32).to(device)
criterion = FocalLossWithSmoothing(
    num_classes=num_classes,
    gamma=2.0,
    alpha=alpha_tensor,
    smoothing=0.1,
    reduction='mean'
)

```

• **Appendix C:** Presents The Train and evaluate DenseNet121 Cell in Kaggle Editor Notebook.

#Train and evaluate DenseNet

```

model_densenet =
models.densenet121(weights=DenseNet121_Weights.IMAGENET1K_V1)
in_features = model_densenet.classifier.in_features
model_densenet.classifier = nn.Sequential(
    nn.Dropout(p=0.3),
    nn.Linear(in_features, num_classes)
)

criterion = FocalLoss(gamma=2.0, alpha=alpha_tensor, reduction='mean')
optimizer = optim.Adam(model_densenet.parameters(), lr=0.0001, weight_decay=1e-
4)

history_densenet = train_model(model_densenet, train_loader, val_loader, criterion,
optimizer, num_epochs=30, model_name='DenseNet')

plt.figure(figsize=(12, 4))
plt.subplot(1, 2, 1)
plt.plot(history_densenet['train_loss'], label='Train Loss')
plt.plot(history_densenet['val_loss'], label='Val Loss')
plt.title('DenseNet Loss')
plt.legend()
plt.subplot(1, 2, 2)
plt.plot(history_densenet['train_acc'], label='Train Acc')
plt.plot(history_densenet['val_acc'], label='Val Acc')
plt.title('DenseNet Accuracy')
plt.legend()
plt.savefig('densenet_training.png')
plt.close()
evaluate_model(model_densenet, test_loader, 'DenseNet', le.classes_)

```

• **Appendix D:** Presents The Train and evaluate DenseNet121 and EfficientNet_B0 Cell in Kaggle Editor Notebook using CrossEntropyLoss.

```
#Train and evaluate DenseNet
```

```
model_densenet = models.densenet121(pretrained=True)
```

```

model_densenet.classifier      =      nn.Linear(model_densenet.classifier.in_features,
num_classes)
criterion = nn.CrossEntropyLoss(weight=class_weights)
optimizer = optim.Adam(model_densenet.parameters(), lr=0.001)
history_densenet = train_model(model_densenet, train_loader, val_loader, criterion,
optimizer, num_epochs=30, model_name='DenseNet')
plt.figure(figsize=(12, 4))
plt.subplot(1, 2, 1)
plt.plot(history_densenet['train_loss'], label='Train Loss')
plt.plot(history_densenet['val_loss'], label='Val Loss')
plt.title('DenseNet Loss')
plt.legend()
plt.subplot(1, 2, 2)
plt.plot(history_densenet['train_acc'], label='Train Acc')
plt.plot(history_densenet['val_acc'], label='Val Acc')
plt.title('DenseNet Accuracy')
plt.legend()
plt.savefig('densenet_training.png')
plt.close()
evaluate_model(model_densenet, test_loader, 'DenseNet', le.classes_).

```

/usr/local/lib/python3.11/dist-packages/torchvision/models/_utils.py:208:

UserWarning: The parameter 'pretrained' is deprecated since 0.13 and may be removed in the future, please use 'weights' instead.

```
warnings.warn(
```

/usr/local/lib/python3.11/dist-packages/torchvision/models/_utils.py:223:

UserWarning: Arguments other than a weight enum or `None` for 'weights' are deprecated since 0.13 and may be removed in the future. The current behavior is equivalent to passing `weights=DenseNet121\_Weights.IMAGENET1K\_V1`. You can also use `weights=DenseNet121\_Weights.DEFAULT` to get the most up-to-date weights.

```
warnings.warn(msg)
```

Downloading: "https://download.pytorch.org/models/densenet121-a639ec97.pth" to /root/.cache/torch/hub/checkpoints/densenet121-a639ec97.pth

100% ██████████ 30.8M/30.8M [00:00<00:00, 156MB/s]

Epoch [1/30] Train Loss: 1.6362, Train Acc: 46.39%, Val Loss: 1.6097, Val Acc: 50.47%

Epoch [2/30] Train Loss: 1.6007, Train Acc: 48.12%, Val Loss: 1.6831, Val Acc: 56.27%

Epoch [3/30] Train Loss: 1.4964, Train Acc: 50.18%, Val Loss: 1.3868, Val Acc: 52.09%

Epoch [4/30] Train Loss: 1.3909, Train Acc: 54.02%, Val Loss: 1.6953, Val Acc: 50.66%

Epoch [5/30] Train Loss: 1.3355, Train Acc: 54.39%, Val Loss: 1.3566, Val Acc: 54.59%

Epoch [6/30] Train Loss: 1.2427, Train Acc: 56.72%, Val Loss: 1.2605, Val Acc: 59.64%

Epoch [7/30] Train Loss: 1.1923, Train Acc: 57.76%, Val Loss: 1.3450, Val Acc: 56.89%

Epoch [8/30] Train Loss: 1.1588, Train Acc: 59.28%, Val Loss: 1.1456, Val Acc: 60.76%

Epoch [9/30] Train Loss: 1.1474, Train Acc: 60.95%, Val Loss: 1.4932, Val Acc: 49.91%

Epoch [10/30] Train Loss: 1.1357, Train Acc: 60.79%, Val Loss: 1.2306, Val Acc: 53.34%

Epoch [11/30] Train Loss: 1.1757, Train Acc: 58.40%, Val Loss: 1.1550, Val Acc: 61.63%

Epoch [12/30] Train Loss: 1.0739, Train Acc: 61.87%, Val Loss: 1.1306, Val Acc: 59.01%

Epoch [13/30] Train Loss: 1.0222, Train Acc: 63.13%, Val Loss: 1.2133, Val Acc: 64.50%

Epoch [14/30] Train Loss: 1.0184, Train Acc: 63.72%, Val Loss: 1.1534, Val Acc: 68.50%

Epoch [15/30] Train Loss: 0.9956, Train Acc: 63.27%, Val Loss: 1.1483, Val Acc: 61.95%

Epoch [16/30] Train Loss: 1.0296, Train Acc: 62.62%, Val Loss: 1.1357, Val Acc: 63.63%

Epoch [17/30] Train Loss: 0.9417, Train Acc: 64.57%, Val Loss: 1.0566, Val Acc: 63.63%

Epoch [18/30] Train Loss: 0.9493, Train Acc: 64.88%, Val Loss: 1.0660, Val Acc: 59.76%

Epoch [19/30] Train Loss: 0.9090, Train Acc: 67.42%, Val Loss: 1.1719, Val Acc: 66.19%

Epoch [20/30] Train Loss: 0.8914, Train Acc: 66.17%, Val Loss: 1.2697, Val Acc: 60.82%

Epoch [21/30] Train Loss: 0.9701, Train Acc: 64.36%, Val Loss: 1.0129, Val Acc: 65.69%

Epoch [22/30] Train Loss: 0.9035, Train Acc: 66.28%, Val Loss: 1.0839, Val Acc: 67.12%

Epoch [23/30] Train Loss: 0.8340, Train Acc: 68.72%, Val Loss: 1.0815, Val Acc: 71.87%

Epoch [24/30] Train Loss: 0.8497, Train Acc: 67.58%, Val Loss: 1.0488, Val Acc: 68.93%

Epoch [25/30] Train Loss: 0.8134, Train Acc: 68.03%, Val Loss: 1.1718, Val Acc: 75.92%

Epoch [26/30] Train Loss: 0.8003, Train Acc: 68.67%, Val Loss: 0.9954, Val Acc: 69.56%

Epoch [27/30] Train Loss: 0.7709, Train Acc: 69.25%, Val Loss: 1.0768, Val Acc: 70.74%

Epoch [28/30] Train Loss: 0.7570, Train Acc: 68.34%, Val Loss: 1.0332, Val Acc: 68.93%

Epoch [29/30] Train Loss: 0.7409, Train Acc: 69.09%, Val Loss: 1.0633, Val Acc: 64.82%

Epoch [30/30] Train Loss: 0.7163, Train Acc: 69.65%, Val Loss: 1.0668, Val Acc: 70.74%

Best model saved to best_densenet.pth

Current working directory: /kaggle/working

Looking for file: best_densenet.pth

Files in directory: ['best_densenet.pth', 'densenet_training.png', '__notebook__.ipynb']

DenseNet Test Accuracy: 72.34%

DenseNet Precision: 0.7947

DenseNet Recall: 0.7234

DenseNet F1 Score: 0.7452

Epoch [1/30] Train Loss: 1.5816, Train Acc: 48.65%, Val Loss: 1.3163, Val Acc: 49.66%

Epoch [2/30] Train Loss: 1.3790, Train Acc: 52.55%, Val Loss: 1.3274, Val Acc: 47.91%

Epoch [3/30] Train Loss: 1.2469, Train Acc: 58.25%, Val Loss: 1.2388, Val Acc: 56.21%

Epoch [4/30] Train Loss: 1.0938, Train Acc: 62.85%, Val Loss: 1.1792, Val Acc: 61.63%

Epoch [5/30] Train Loss: 1.0885, Train Acc: 63.05%, Val Loss: 1.1039, Val Acc: 66.81%

Epoch [6/30] Train Loss: 0.9793, Train Acc: 64.39%, Val Loss: 1.0373, Val Acc: 56.21%

Epoch [7/30] Train Loss: 0.9340, Train Acc: 65.60%, Val Loss: 1.0023, Val Acc: 57.89%

Epoch [8/30] Train Loss: 0.8697, Train Acc: 66.66%, Val Loss: 0.9217, Val Acc: 70.93%

Epoch [9/30] Train Loss: 0.8002, Train Acc: 67.39%, Val Loss: 1.7874, Val Acc: 58.64%

Epoch [10/30] Train Loss: 0.7613, Train Acc: 68.86%, Val Loss: 0.8799, Val Acc: 68.75%

Epoch [11/30] Train Loss: 0.7893, Train Acc: 69.81%, Val Loss: 1.0028, Val Acc: 57.33%

Epoch [12/30] Train Loss: 0.7027, Train Acc: 70.99%, Val Loss: 0.9945, Val Acc: 66.00%

Epoch [13/30] Train Loss: 0.7174, Train Acc: 68.62%, Val Loss: 1.0460, Val Acc: 69.43%

Epoch [14/30] Train Loss: 0.6850, Train Acc: 69.85%, Val Loss: 0.8148, Val Acc: 74.67%

Epoch [15/30] Train Loss: 0.6013, Train Acc: 73.27%, Val Loss: 0.9884, Val Acc: 68.68%

Epoch [16/30] Train Loss: 0.5980, Train Acc: 72.88%, Val Loss: 1.0978, Val Acc: 68.93%

Epoch [17/30] Train Loss: 0.5345, Train Acc: 74.15%, Val Loss: 0.8805, Val Acc: 73.61%

Epoch [18/30] Train Loss: 0.5370, Train Acc: 74.66%, Val Loss: 1.2814, Val Acc: 69.81%

Epoch [19/30] Train Loss: 0.5539, Train Acc: 74.55%, Val Loss: 0.9229, Val Acc: 72.18%

Epoch [20/30] Train Loss: 0.4600, Train Acc: 76.22%, Val Loss: 0.8940, Val Acc: 74.24%

Epoch [21/30] Train Loss: 0.5879, Train Acc: 73.19%, Val Loss: 0.8692, Val Acc: 68.81%

Epoch [22/30] Train Loss: 0.5079, Train Acc: 74.97%, Val Loss: 0.9935, Val Acc: 74.30%

Epoch [23/30] Train Loss: 0.5541, Train Acc: 75.07%, Val Loss: 1.0705, Val Acc: 67.12%

Epoch [24/30] Train Loss: 0.5561, Train Acc: 75.28%, Val Loss: 0.8467, Val Acc: 71.80%

Epoch [25/30] Train Loss: 0.4021, Train Acc: 79.14%, Val Loss: 0.9697, Val Acc: 73.74%

Epoch [26/30] Train Loss: 0.4149, Train Acc: 77.56%, Val Loss: 0.8828, Val Acc: 72.05%

Epoch [27/30] Train Loss: 0.4621, Train Acc: 77.30%, Val Loss: 1.1310, Val Acc: 76.29%

Epoch [28/30] Train Loss: 0.3570, Train Acc: 79.48%, Val Loss: 1.0614, Val Acc: 80.10%

Epoch [29/30] Train Loss: 0.3509, Train Acc: 80.92%, Val Loss: 0.9874, Val Acc: 74.61%

Epoch [30/30] Train Loss: 0.3623, Train Acc: 80.00%, Val Loss: 1.1149, Val Acc: 74.73%

Best model saved to best_efficientnet.pth

Current working directory: /kaggle/working

Looking for file: best_efficientnet.pth

Files in directory: ['best_efficientnet.pth', 'DenseNet_confusion_matrix.png', 'best_densenet.pth', 'densenet_training.png', 'efficientnet_training.png', '__notebook__.ipynb']

EfficientNet Test Accuracy: 74.24%

EfficientNet Precision: 0.8104

EfficientNet Recall: 0.7424

EfficientNet F1 Score: 0.7621

• **Appendix E:** Presents The Max-Voting Ensemble Results when using CrossEntropyLoss and the Figure of Confusion-Matrix Max-Voting Ensemble Results.

Max-Voting Ensemble Results when using CrossEntropyLoss

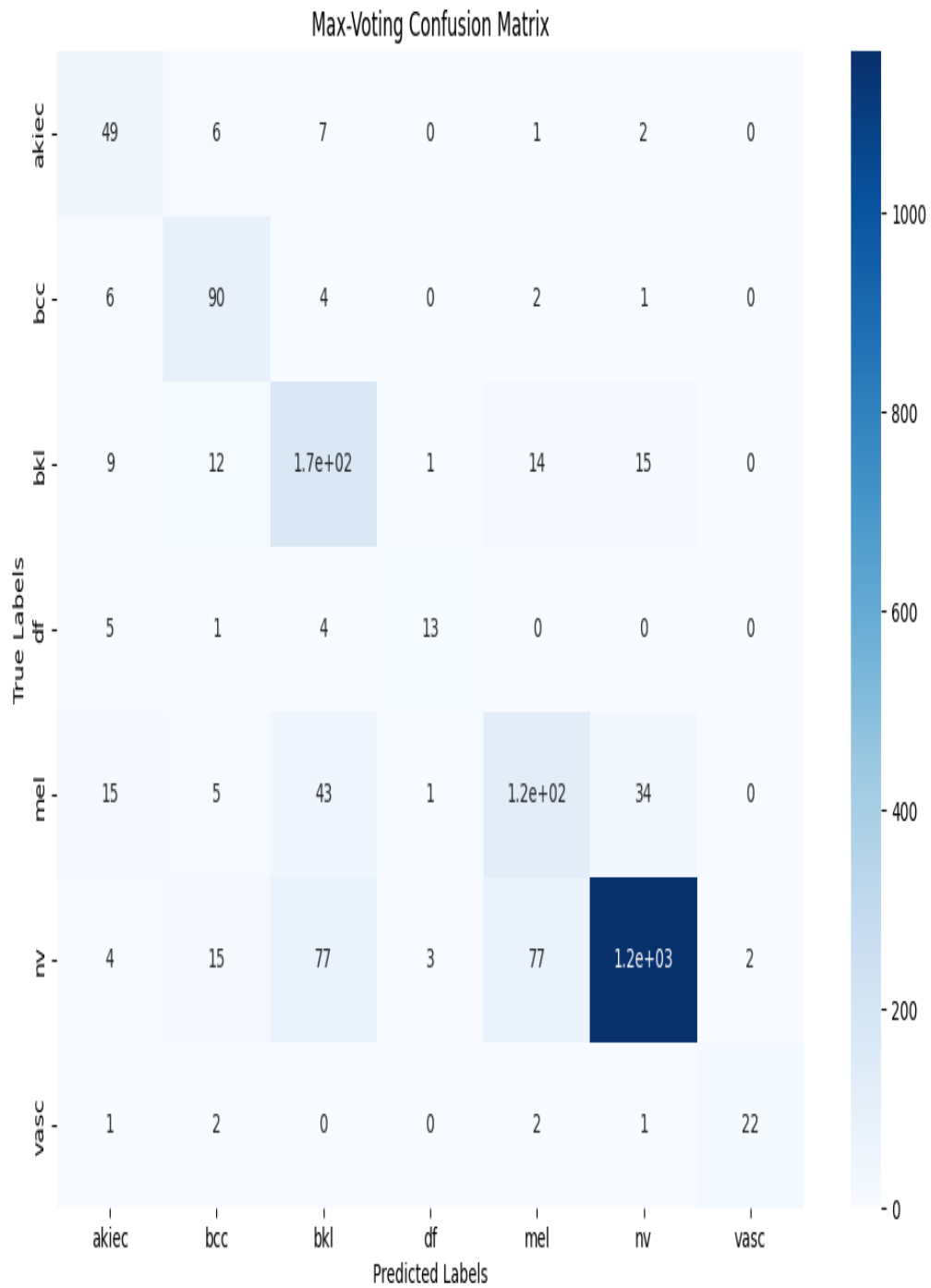
	precision	recall	f1-score	support
akiec	0.55	0.75	0.64	65
bcc	0.69	0.87	0.77	103
bkl	0.56	0.77	0.65	220
df	0.72	0.57	0.63	23
mel	0.57	0.56	0.56	223
nv	0.96	0.87	0.91	1341
vasc	0.92	0.79	0.85	28
accuracy			0.81	2003
macro avg	0.71	0.74	0.71	2003
weighted avg	0.84	0.81	0.82	2003

Accuracy: 0.8142785821268098

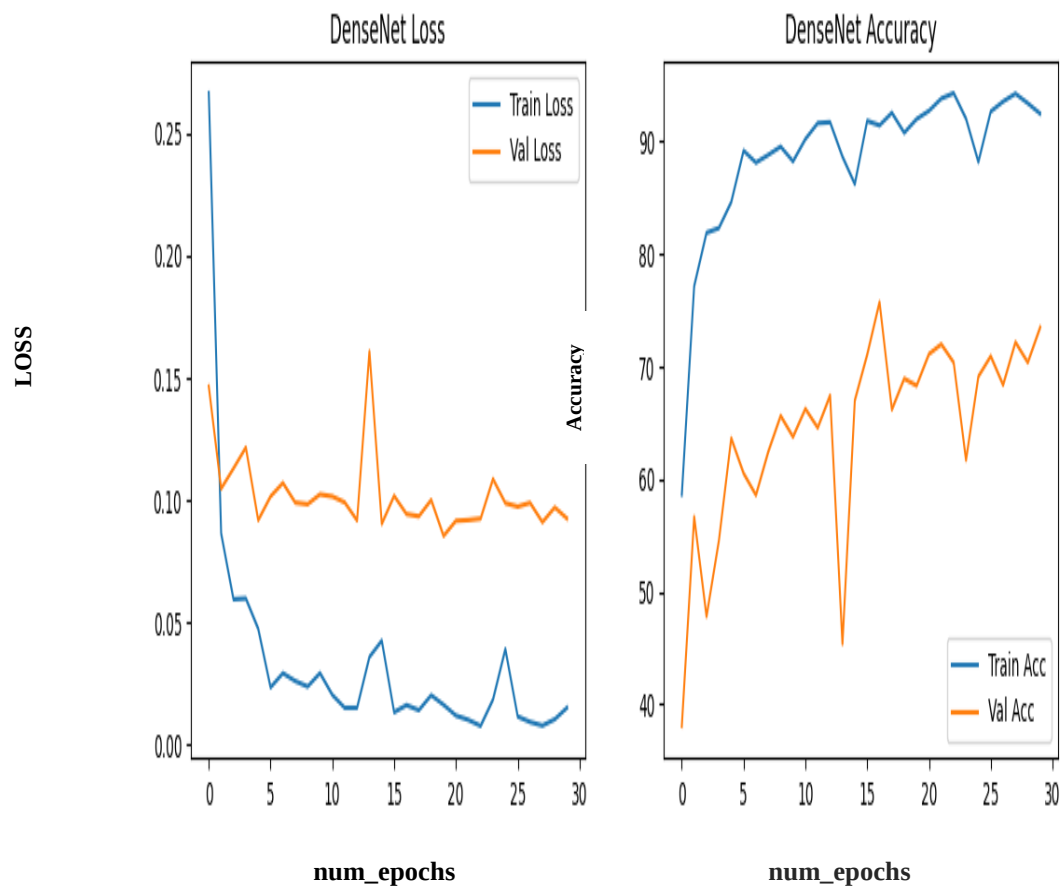
Precision (macro): 0.7077742819213363

Recall (macro): 0.7392210585423253

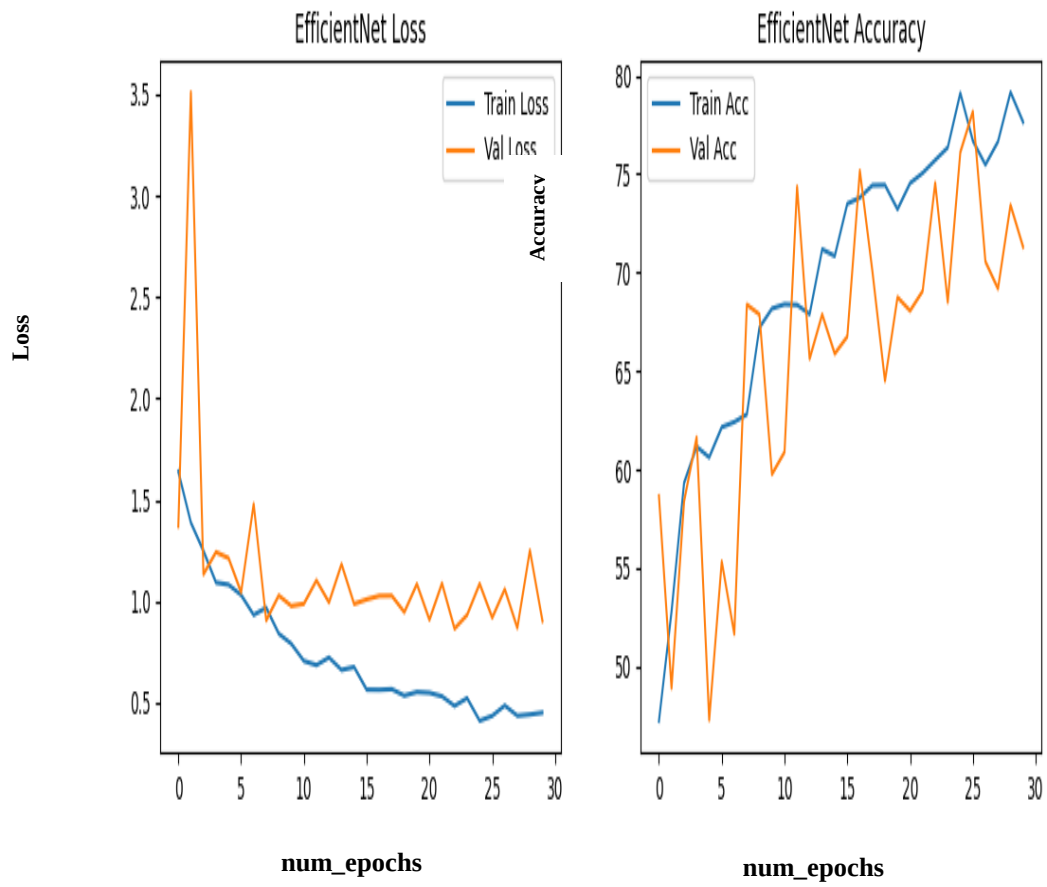
F1 Score (macro): 0.7148079402488591



Figure(A-1): Confusion-Matrix Max-Voting Ensemble Results when using CrossEntropyLoss.



Figure(A-2): Detailed Analysis of DenseNet121 Training and Validation Performance Over 30 Epochs when using CrossEntropyLoss.



Figure(A-3): Detailed Analysis of EfficientNet-B0 Training and Validation Performance Over 30 Epochs when using CrossEntropyLoss.