Real-Time Viral Skin Lesion Diagnosis: A Hybrid Deep Learning Framework for On-Device Deployment

BY

Md. Sakibuzzaman Alif
ID: 242-25-042

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Masters of Science in Computer Science and Engineering

Supervised By

Ms. Nazmun Nessa Moon
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2025

APPROVAL

This Thesis titled **"Real-Time Viral Skin Leison Diagnosis: A Hybrid Deep Learning Framework for On-Device Deployment"**, submitted by **Md. Sakibuzzaman Alif**, ID No: **242-25-042** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2025.

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

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

Internal Examiner

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Mr. Nazibur Rahman
Head of IT Infrastructure
Networld Bangladesh PLC

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Ms. Nazmun Nessa Moon, Associate Professor, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Ms. Nazmun Nessa Moon
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Md. Sakibuzzaman Alif
ID: 242-25-042
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartfelt thanks and gratitude to Almighty Allah for His divine blessing, which makes it possible to complete the final year project successfully.

I am grateful and wish to express my profound indebtedness to **Ms. Nazmun Nessa Moon, Associate Professor, Department of CSE,** Daffodil International University, Dhaka, deep knowledge & keen interest in the field of Machine Learning to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to **Dr. Sheak Rashed Haider Noori**, **Head of the Department of CSE**, for his kind assistance in completing our project, as well as to the other faculty members and staff of the CSE department at Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

This thesis addresses a critical public health issue by presenting a hybrid deep learning pipeline for the real-time, on-device diagnosis of viral skin lesions. The core objective was to develop a model that effectively balances high classification performance with low computational cost, enabling its deployment on mobile devices for use in resource-limited environments. A comprehensive comparative study was conducted on five hybrid architectures, each combining a custom-trained Convolutional Neural Network with a powerful pre-trained backbone. Through a rigorous two-staged fine-tuning approach, the Custom CNN + EfficientNetB0 architecture was identified as the most effective, achieving an outstanding classification accuracy of 99%. The selected model was then efficiently quantized into a lightweight 5.6 MB TFLite format, demonstrating a remarkable average on-device inference time of 50 ms. This achievement culminates in the implementation of a high-performing, privacy-preserving, and low-cost model within a functional mobile application. This work underscores the feasibility of developing practical, end-to-end AI diagnostic tools that can support clinical practice and provides a scalable platform for future research in accessible visual-based diagnostic solutions.

Keywords: Viral Skin Lesions, Deep Learning, Hybrid Models, Mobile Application, Medical Image Analysis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Machine diagnosis of dermatology problems in recent years has begun to be transformed with deep learning. Precise discrimination among types of skin infections including virus is needed. Harmful exanthems represent a major infectious disease risk to the population, since they comprise highly contagious viruses: smallpox virus (variola), VZV (chickenpox) and the rubeola virus (measles). Because the approaches focus on screening at huge scale, early accurate diagnosis prompts timely initiation of the interventions of isolation and treatment targeting mechanisms that lead to prevention in community-wide high-incidence settings (including places where such expertise does not yet exist).

Deep learning models (like CNNs) have been successful in skin lesions classification and even seems to be competitive or superior than the human experts on benchmarks [1]. But their feasibility has been a problem in real-world scenarios. The such complex models are also resource expensive and these powerful models cannot be run on the resource constraint smart phones [15]. This limitation also applies to the stiffness of developing user-friendly, transferable and scalable tools that are usable for a variety of stakeholders (besides HCWs in different settings)–a difficulty solved with papers using smartphone-captured photographs [16].

This thesis takes a step in the right direction to address this important gap by recipient side attention development and evaluation of hybrid deep learning models. Following [2, 19] as a recent successful precedent of combining architectures, we construct models to adress the gap between our single structure andéffthem in terms of offering better inheritance properties from classical instancesfortraditionalCNNs (FCIS when discussing local feature extraction) tomodulernapproaches such as Vision Transformers which are revolving around global context[1]. This work also should be enough to allow Coup to have fidelity and actuate. This

work results in a prototype mobile app, and the experiment results on which already prove that our approach can be directly used for on-device automatic and real-time diagnosis.

1.2 Motivation

This study was prompted by the need for a quick and non-invasive diagnosis of viral skin eruptions. Time is of the essence in the clinical setting, and rapid and accurate diagnosis is critical for disease management and for patients. Misdiagnosis is one of the major problems faced in resource-limited areas and could result not only in the waste of resources but also in the uncontrolled spread of the communicable disease in a community. This is a significant challenge in many healthcare practices (both the rural or low-resource settings) where a trained dermatologist or expert opinion may not be easily accessible. This leads to the fact that the number of specialist is limited, resulting in the bottleneck in the healthcare system and the delay in diagnosis and intervention.

Although recent breakthroughs in deep learning have shown great hope for medical image analysis and may potentially provide a solution to this diagnostic bottleneck, a crucial problem exists. Most cutting-edge models, which are better-trained on large datasets and include millions or even billions of parameters, are computationally intensive and challenging for edge devices such as smartphones to run. Such models need to be run on the on-device (local) core, which does not have sufficient processing power, memory and power supply. This computational cost makes a large gap between the success in the research side and application of it.

To overcome this key gap, the current project will pursue innovative diagnostic PARADIGM. The goal is to develop deep learning models that achieve not only high accuracy to predict diagnostic data with precision, but are also efficient and trained on sufficiently general data so that they could potentially be deployed as assistive tools for use by medical clinicians around the world. By considering performance and efficiency together, this study aims to develop a solution that is clinically effective and operationally compelling, democratizing quality diagnostic support.

1.3 Research Objectives

Based on the problems outlined above, the objectives of this research are:

- To conduct a comprehensive comparative analysis of five distinct hybrid deep learning architectures for the multi-class classification of viral skin lesions.
- To identify the optimal hybrid model that provides the best trade-off between classification performance and computational efficiency.
- To convert the best-performing model into a lightweight format (e.g., TFLite) for on-device inference.
- To develop a mobile application that uses the TFLite model to demonstrate the practical, real-time diagnostic capability of the research.
- To produce a comprehensive thesis report detailing the methodology, results, and implications of the project.

1.4 Research Questions

This thesis will seek to answer the following research questions:

- How do hybrid deep learning models, which combine different architectural backbones, perform in comparison to single-backbone models for the task of viral skin lesion classification?
- Which of the five proposed hybrid architectures yields the highest diagnostic accuracy?
- Can the most accurate model be successfully optimized for and deployed on a mobile device to provide real-time diagnostic assistance?
- How does the performance of the hybrid models differ across various lesion classes, particularly for those that are visually ambiguous or have limited data?

1.5 Expected Output

The expected outputs of this research project are:

- A thorough literature review of the state-of-the-art in deep learning for skin lesion classification.
- Five trained hybrid deep learning models, each with a detailed performance analysis.
- A finalized, lightweight TFLite model of the best-performing architecture.
- A functional mobile application capable of performing real-time classification of viral skin lesions.
- A completed thesis report that documents the entire research process, from methodology to conclusions.

1.6 Project Management and Finance

The research work doesn't get fund from any individuals or organization.

1.7 Report Layout

This thesis is organized into six chapters:

- **Chapter 1 (Introduction):** Provides a background of the research, defines the motivation, and outlines the research questions and objectives.
- **Chapter 2 (Background):** Presents a literature review of relevant works in the field and identifies the research gap.
- **Chapter 3 (Research Methodology):** Details the dataset, preprocessing techniques, the five hybrid architectures, and the training and evaluation procedures.
- **Chapter 4 (Experimental Results and Discussion):** Presents the quantitative and qualitative results, including performance metrics, accuracy/loss curves, and confusion matrices.

- **Chapter 5 (Impact on Society, Environment and Sustainability):** Discusses the broader implications and ethical considerations of the work.
- **Chapter 6 (Conclusion and Future Work):** Summarizes the study's findings, provides concluding remarks, and suggests directions for further research.

CHAPTER 2
BACKGROUND

2.1 Introduction

Automated skin lesion classification research has progressed rapidly in the past few years, and this may result from three main reasons: (i) The maturity of deep learning, (ii) A series of large and open public datasets are constructed; (iii) More effective training strategy is able to be proposed. On the other hand, classical techniques are built upon manually designed features and traditional machine learning classifiers, while modern algorithms take advantage of deep neural networks that can automatically learn a cascaded feature representation from raw image data and demonstrate unsurpassed diagnostic accuracy.

For all the advances then, the road from a high-performance research model to a dependable, clinical tool remains impeded by barriers. There are three main obstacles hampering the universal incorporation into the dermatologist's clinic. The first phenomenology is dataset bias and generalizability: models trained on a single homogenous dataset do not retain their performance when deployed over new, test clinical environments, including those with different patient populations, skin types, cameras etc. This is called the domain shift, which undermines the confidence of the model in clinical practice and restricts its practicality.

The second difficulty is computational bound. The state-of-the-art deep learning models, in particular transformer-based or hybrid architectures, are computational expensive and require large memory resource. However, the large size, and high latency of the models makes it challenging to deploy it in real-time, or on edge (resource constrained) devices such as mobile phones or in hand (portable) dermatosopes.

The third, possibly the most important, obstacle is the interpretability. The decision-making of an AI-based diagnostic tool should be transparent and understandable for a clinician to rely on and to use the tool effectively. A “black box” model, no matter how accurate, generates no

justifiable reason for clinical application and meets insurmountable obstacles toward regulatory clearance.

In this chapter, we survey these works, which cover various architectural paradigms including convolutional networks (CNNs) and transformers to both complex hybrid models and efficiency focus models. We will emphasize common trends, areas of frequent limitation, point out open problems for further exploration. I specifically look into works with the HAM10000, ISIC and their derivatives, as they define the backbone of the literature and therefore dictate the direction of the field. Finally, we advocate the primacy importance of the MCVSL dataset in its consideration as a more representative and valuable auxiliary dataset to verify generalization as well as fairness, seeking to more directly support the development of robust and equitable AI systems for dermatology [21].

2.2 Related Works

2.2.1 Foundations and Early Innovations

The pre-deep learning time of skin lesion analysis was characterized by old-fashioned supervised pipelines based on hand-engineered features. This two-step step process requires the extraction of simple numerical descriptors of the image, typically color histograms (e.g., RGB or CIELAB mean and standard deviation), texture (e.g., GLCM), and shape (circularity, eccentricity, aspect ratio). These feature vectors were subsequently input to standard classifiers such as Support Vector Machines (SVM) or Random Forest. This method was time-consuming and often led to fragile models that were heavily conditioned on the quality of the feature engineering.

Convolutional Neural Networks (CNNs) have revolutionized the concept of end-to-end learning of features. CNNs learn to represent the image data in a hierarchical way, with lower level features (e.g., edges, textures) represented at the early layers of the network and high level features (e.g., object) appear at the deep layers of the network. This end-to-end method

essentially replaced manual feature engineering and set significantly stronger benchmarks on benchmark datasets.

The CNN baselines were constructed with the AlexNet, VGGNet and ResNet architecture with separate parameterization and fine-tuning, and to be compared against a large number of available methods. Gupta et al. obtained a validation accuracy of ~ 88% on HAM10000 with a custom CNN, emphasising the critical importance of regularization and addressing class imbalance - the latter being an affliction in medical datasets, where certain conditions may have orders of magnitude frequency difference [16]. Multi-backbone comparisons with established architectures such as VGG16, DenseNet and ResNet also reported mixed though competitive results (85% accuracy), but performance showed a high dependency on the choice of training split (and subsequent augmentations strategies) [12]. That variability hammered home an important point: the model is only part of it; strategy around training and quality of data are just as if not more important.

But there was still worth in the old-fashioned pre-processing as applied to deep learning, according to a few academics. For example, Pandey et al. concatenated non-local means denoising and sparse dictionary features before CNN stage and approximately 90% of the classification accuracy [17].

They found that their method performed satisfactorily and was sensitive to noisy input or very small lesions. Likewise, Jabber et al. used cluster K-means segmenting in CIELAB color to delineate the accurate area of lesions, and as the attentional cue for CNN. The system obtained an accuracy of 87% with good quality results for lesions that have indistinct borders (low contrast, low gradient), and inferior results on small or low-contrast lesions [3]. These early studies collectively imply two primary learnings: (i) simple CNNs are incredibly capable, but also ridiculously sensitive to training protocol and extremely class-imbalanced data -- both big practical issues; (ii) classical pre-processing can be very helpful in a few corner cases; however they introduce another dangerous form of brittleness into the pipeline making it at the end yet critically sensitive to real world data [13, 12].

2.2.2 Hybrid and Attention-Based Architectures

In order to overcome the limitations of vanilla CNNs in capturing the global context, as they are limited to learn context with a fixed receptive field, a recently emerging line of work introduced explicit mechanisms to model long-range dependencies. This paved the way to the hybrid models that combine the strengths of the CNNs with the expressive power of the transformer blocks or attention mechanisms.

CNN + Transformer hybrids have been most promising. Transformers for NLP were designed to efficiently capture long distance dependencies in a sequence of tokens. In computer vision, a model might break an image into patches, or "tokenize" it. Gulzar et al. achieved an outstanding 97.57% validation accuracy on HAM10000 by applying a small CNN to encode a set of feature vectors that serve as tokens along which a transformer layer that computes self-attention can operate [1]. This method played to the strengths of a CNN to learn local, textural features, as well as the ability of a transformer to contextualize global information across different regions of an image. Similarly, Kumar et al. added one transformer layer on top of an existing CNN backbone for ISIC 2019 with 96% of accuracy and contributed the improvement to the ability of transformer to capture the long-range dependencies, with trade off of increased training cost [14]. Mohan et al. also followed a transformer-dominant pipeline to achieve around 94% accuracy on HAM10000 but argued the huge edge-device limitation for these models as hard to deploy in clinical scene specifically [19]. Other works, like Wang et al., investigated Transformer-CNN but without performing head-to-head accuracy comparison, emphasizing the necessity of standard reporting and cross-dataset evaluation to justify these architectural choices [8].

With respect to explicit attention mechanisms, the latter are more fine-grained. CitesAkbari et al Non-local neural networksYeah but while I work with LM, the hierarchical attention mechs that some people are aware of end up doing what a transformer does over conv layers effectively and is also much simpler (& faster) than one based on a transformer. Instead of taking full transformer block, they take a module which only takes weighted sum of the features - Model can learn what part of image is actually relevant. Reddy et al. included an attention

module into their CNN-Transformer pipeline and reported 92.4% accuracy, and suggested that the visualized attention maps could be help full for clinical reasoning [4]. This led to an explicit interpretability with the price of additional inference time. Agarwal & Mahto come up with an even more complex fusion stack of CNNs, Transformers, and a Knowledge-Attentive Network (KAN) to achieve c.a 92% on ISIC 2018 but mention how little was known about which part did what for explainability —a common thread in many hybrids [2].

Last but not least, purely vision transformer models that dispense with CNN backbone all together and consider self-attention only have also been transformed to dermoscopy. Himel et al. employed a Vision Transformer (ViT) model on ISIC 2018 with an outstanding accuracy of around 96.1% [20]. This example showed the effectiveness of self-attention for dermoscopic images, but they also recognized that huge computational resources and collection of a large number of dermoscopic images are required for training those models from scratch. Modern CNNs and hybrid models are still under development, with Zhang et al. investigating the use of AgingCNN, a lightweight CNN, in a hybrid pipeline [6]. The non-reporting of full metrics in their study serves as an important reminder that partial evaluation can hide a model's generalization limits for state-of-the-art backbones.

Overall, the literature suggests that attention and hybridization can lead to a better accuracy compared plain CNNs, especially when lesions manifest with diffuse boundaries or images are corrupted by background artifacts. But such gains frequently come with suboptimal training and runtime efficiency as well as a lack of interpretability on the reasons why the model works the way it does, which thus requires further study for efficient and interpretable designs.

2.2.3 Multimodal, Foundation, and Lightweight Solutions

Work that is closely related to ours, but is more focused on practical, deployment-oriented solutions, which are efficient, robust to the scarce label situation, and employ diverse supervision sources. These methods acknowledge that clinical utility is determined not only by peak accuracy, but also by more comprehensive considerations such as computational cost and label-efficiency.

Efficient and light-weight CNNs is a cornerstone of this methodology. EfficientNet model family, which is based on the idea of compound scaling provides the best trade-off between accuracy and computation costs. As reported in AIP Advances, EfficientNet-B3 only achieved approximate 84% accuracy on HAM10000 and it had to be carefully tuned to cope with the multi-class imbalance and to give an acceptable performance on minority classes [5]. Tejasri et al. showed that strong data augmentation and transfer learning enable them to achieve approx (91 %) when trained jointly on ISIC and HAM10000, implying that these well-compounded backbones still remain competitive baselines [13]. ” Transfer learning is of particular utility in dermatology, as features learned on a large dataset such as ImageNet offer a strong initialisation for the more domain-specific skin lesion classification task which in practice might be hindered by a low number of labeled example.

Ensembles and structured learners offer an alternative method of improving performance by combining the proprellors of many models. Hasan & Rifat extracted CNN-based features and used a tree-based learner, such as Gradient-Boosted Decision Tree (GBDT), for learning, achieving roughly 94% accuracy on ISIC 2019 (with use of HAM10000 dataset mentioned) [9]. This also leads to heavy computational and memory load during training. Agarwal et al. also suffered from severe class imbalance, with an accuracy of 91.8% on HAM10000, calling for better loss design and sampling strategies to achieve better performance balance [15].

The high cost of expert annotation in medical imaging has driven heavy interest in semi-supervised learning and knowledge distillation. Manivannan employed an semi-supervised approach including knowledge distillation(KD), and achieved on ISIC with AUC $\approx$ 0.91 [10]. This showed a great potential to enhance label-efficiency, especially for clinics that have no luxury of time labeled. In this approach, a large model “teacher” (which is powerful and memory-intensive) is trained on a small set of labeled data and a huge pile of unlabeled images. The teacher then generates "soft labels" for the unlabeled data, and is used to train a smaller (and more efficient) "student" model. This method has the advantage of effectively transferring/dispatch the knowledge from the heavyweight teacher to its lighter student.

The revival of pre-processing and fusion techniques also indicates a(practical trend. Akter et al. proposed a hybrid feature fusion method on CNNs and obtained 92.27% accuracy [11], Jabber et al. s segmentation-tasked pipeline [3] demonstrates that there is still place and benefit for hand-designed steps to work along deep features. However, the robustness of these methods needs to be well verified, so that they have not brittleness when used for a new imaging condition. However, despite such distinctions in these different threads, the literature seems to strike an overall balance of three fundamental (desirable) properties: summarizes that no 'one size fits all' family is prevalent under every constraint.

2.3 Key Challenges in Skin Lesion Classification

The problem of how to put these models working in realistic field (i.e., low-resources) applications is however a major issue; even a very performing model is not worth if not usable in median devices. The latter has also resulted in a second line of research interests related to light-weight models that compromise diagnostic performance versus the computational budget. This trade-off was also well-captured by a model for mobile and embedded vision applications, namely MobileNetV3 [6]. We have demonstrated through this study an on-device test potential as rapid POC testing at moderate internet penetration sites [10]. Databases as the PAD-UFES20, which shows real data for smartphone images are a good attempt to make models aware of the characteristics of real data and that they could change [16]. More complex training setting are also investigated for improving the model's performance and generalization. For example, curriculum learning, a multimodal model used to obtain the highest accuracy of 96.5%, was obtained by showing a model progressively more complex examples over time Evening out curriculum which is expected to mimic human learning [11]. Another model which used self-supervised pretraining and domain adaptation in training for large scale unlabeled data then fine-tunning the high accuracy and generalization able to achieved was demonstrated [12]. Motivating the study into uncertainty-aware models that use methods such as Monte Carlo dropout [13] is the dependence for model interpretability and trust in high-stake medical environments too. These models not only provide a classification but provide some indication of the confidence in the prediction, which is critical for a clinician

at that last step. Multi-stage over the whole training pipeline of end-to-end models has also been investigated, a cascade model has been designed for a two-stages training where in the first stage the localization of the lesion was performed using a UNet segmentation, followed by cutting the region containing the lesion and training a custom CNN-based model classifier [9]. This two-step system achieves overall 97.3% accuracy and demonstrates that combining multiple simple computer vision tasks can provide a more robust and informative diagnostic system.

2.3.1 Dataset Limitations and Generalizability

A pervasive issue in the field is the heavy reliance on a small number of benchmark datasets, primarily HAM10000 and specific ISIC editions. While these datasets have been instrumental in advancing the field, they are not representative of the real-world patient population. They differ significantly in their imaging protocols, lesion distributions, and, most critically, the diversity of skin tones represented. This lack of diversity leads to a fundamental problem: a model that performs exceptionally well on its training dataset often suffers a significant drop in performance when encountering data from a different distribution—a phenomenon known as domain shift. This happens because the model learns spurious correlations specific to the training data, such as a particular imaging device's artifacts or the lighting conditions of a single clinic, which do not hold true in new environments.

- **Manifestations of the Problem:** Several works points either explicitly or implicitly to this need, due to their lack of cross-dataset testing [6–8, 12, 15, 18]. Khan et al. and Agarwal et al. both stress that their model performance is particularly class-unbalanced and training cohort-specific sensitive [7, 15]. For example, if a dataset is over-represented by frequent benign lesions from a particular clinic, a model might associate certain lighting conditions with a benign diagnosis and be unsuccessful in classifying the same lesion type of a different clinic. Gupta et al. and Pradeepa et al. also demonstrate that picking a backbone architecture can, itself, provide inconsistent gains when used with different ways to split training data and augment data [12, 16]. Furthermore, preprocessing-intensive techniques, when effective on a dataset, may be

fragile to different types of noise or acquisition conditions [3, 17]. To the contrary, this fragility decreases a clinician's confidence in the model to transfer beyond a limited well-maintained setting.

- **Why MCVSL matters**: In order to address the above issues more directly, I introduce the MCVSL (Mendeley) dataset as an alternative, and more diverse, fine-grained benchmark [21]. While HAM10000, as mentioned above, is heavily biased towards light skin phototypes, MCVSL consists of a diverse set not only in terms of skin but also lesion types, hence it can be used as powerful tool in testing transferability and fairness. With the addition of MCVSL to HAM10000/ISIC, cross-dataset validation is feasible, i.e., with a model trained in one of the cohorts and tested in the other.

2.3.2 Computational and Deployment Constraints

The trend of achieving higher accuracy increases the complexity and computational burden of proposed architectures and is difficult to be practically deployed. Transformer layers and hybrid stacks, while very performant, also significantly increase memory needs and inference processing time, which make these approaches unacceptable for real-time or edge deployment on resource-limited hardware such as mobile phones or evening embedded systems in clinics.

The Trade-off between Power and Efficiency: The computational expense of these procedures is likely to be considerable. A huge transformer model possibly uses dozens of GB of vram at infer time and already might take seconds to process a single image, while a toy-size CNN works on few hundreds MB with ms latency. For examples, the measurements of Reddy et al. and Kumar et al. recognize the eciency problem of their attention/transformer blocks despite the high rigid accuracy [4, 14]. Mohan et al. Specifically, flag edge-deployment constraints for a transformer-heavy architecture [19] (to the best of our knowledge, Trott et al. also observe the high training and inference cost of pure-ViT models [20]. Ensemble models, like the CNN + GBDT pipeline, may also be compute-hungry when training and need substantial model compression and optimization during deployment in order to be used effectively [9]. In this setting, lightweight architectures (i.e. EfficientNet-B3) and transfer

learning provide a more practical alternative by lowering computational overheads, at the risk of loss in performance on minority classes unless carefully balanced and calibrated [5, 13, 18].

The Promise of Distillation: Semi-supervised and knowledge distillation workflows provide a good trade-off: in both, the heavy computation is done at training time so that the cost of prediction is light and fast [10]. You can do this by training a large, powerful model, called a teacher, and then using its outputs to train a much smaller, more efficient, student. By imitation, the student is led to model the teacher, which in turn allows knowledge to be compressed to something that is more easily deployable. It offers an efficient route to high accuracy without the associated runtime expenses, and is therefore well suited for embedding in clinical devices with constrained computational capacity.

2.3.3 Interpretability and Clinical Trust

In order for a diagnostic AI system to be accepted by clinicians, it cannot be only a very accurate "black box". It needs to offer reasons for its predictions that are clear to humans. Practitioners must be able to understand why a model makes a certain decision, that is, the explanation is vital to develop trust and is sometimes required by regulations. For instance, FDA guidance on AI in medical devices stress the importance of transparency and that a clear understanding of the model's performance across various scenarios.

The Explainability Gap In this context there are already viable approaches which create understandable outputs, however in many works they are either not used, or if used only utilized in a limited manner. The explainability gap is captured in Agarwal & Mahto's complex fusion architecture [2]. They stress the challenge of assigning a specific prediction to one particular part of a multi-part pipeline. Agarwal et al. also note that calibration problems and lack of performance on imbalanced classes can decrease a clinician's trust in multi-class outputs of a model [15]. Despite the fact that visual interpretability approaches such as saliency maps and Grad-CAM can identify crucial areas, they are not always sufficient. A model could become obsessed with a dermoscope's border, a ruler in the image or a hair follicle, and throw a spurious correlation that happened to produce a "correct" answer but for the incorrect reason.

This can result in a clinician reaching the wrong conclusion, or losing faith in the model altogether.

Clinical Relevance: Integration techniques (including segmentation) Such as the method of Jabber et al. provide a better focus on the abnormalities, as it could be also misleading in noysesimilar images [3]. Headline-Attention-Capturing AON models, such as Reddy et al. ‘s, for example they are optimistic in their sense that they can display image based evidence supporting a prediction [4]. However, for these highlighted regions to become clinically viable, they need to match known skin features –by example, those defining the ABCDE criteria of melanoma detection (Asymmetry, Border, Color, Diameter, Evolution) [2] or any other strong dermoscopic subjective sign. Akter et al. and Tejasri et al. emphasise that any model, with or without fusion and also consisting of only a backbone architecture, needs to be supported by wellcalibrated probabilities as well as clinically interpretable evidence in order for it to truly be actionable and trust worthy [11, 13]. There is also a need to quantify a model’s uncertainty, so it can “abstain” from giving a diagnosis when it is not confident enough, which is a safer approach in a high-stakes clinical environment.

2.4 Research Gap

While the literature reviewed provides a strong evidence of the efficacy of deep learning deep and hybrid models in the classification of skin lesions and other clinical image there is an obvious research void that requires this solution to be attempted. Among the frequently mentioned deficiencies, there is an absence of a large-scale comprehensive comparison that involves all released newly developed fusion algorithms in the literature. The fragmented nature of this research landscape makes it very difficult to nail down general conclusions as to which architectural paradigms in general (an cnn, a chain of cnns or a cnn with Transformers for instance) work better than others. However, a comprehensive study is still missing that would systematically evaluate multiple sets of CNN-CNN and CNN-Transformer models on a single dataset to determine the optimal balance of good performance and computational efficiency.

Lastly, although practical deployable models are considered highly desirable in the community, few works show a complete end-to-end solution ranging from model object detection networks to the optimization for on- device deployment and then integration into a mobile application. Yet the vast majority of research goes only to the point where the model does well on a benchmark and fails to even consider the engineering challenges of getting that model into the hands of a front-line healthcare worker.

This unmet need is addressed in this thesis with a full head-to-head comparison of 5 hybrid models for viral skin lesion classification. By looking at both architectures on the same dataset and indicating performance, we hope to give clear based advice of which way to go. A systematic comparison will help to provide insights into the trade-offs between accuracy, inference speed and model size in architecture design. The culmination of the development of this system into a fully implemented app, results in an end-to-end platform that combines theoretical enquiry to clinical standard. In this sense, our work serves as a pragmatic and standardized preparation for AI-supported dermatology and an instant show piece of efficient diagnosing tools in practice.

Table 2.1: Research Metrics

Sl No	Paper	Dataset Used	Model	Accuracy	Key Gaps/Limitations
[1]	Gulzar et al. (2025)	HAM10000	Hybrid CNN + Transformer	97.57% (Validation)	Needs more diverse datasets beyond HAM10000 to generalize performance.
[2]	Agarwal & Mahto (2025)	ISIC 2018, HAM10000	CNN + Transformer + KAN Fusion	92% (ISIC)	Limited explanation of fusion method's impact on explainability.
[3]	Jabber et al. (2025)	HAM10000	CNN with K-means + LAB-color segmentation	87%	Could improve performance on small or noisy datasets.
[4]	Reddy et al. (2024)	HAM10000	Attention-CNN-Transformer	92.4%	Needs better computational efficiency for

					deployment in real-time settings.
[5]	AIP Advances (2025)	HAM10000	EfficientNet-B3	84%	Model needs optimization for multi-class skin cancer classification.
[6]	Zhang et al. (2025)	HAM10000	ConvNeXt + CNN	N/A	Performance metrics not explicitly reported; need more evaluation on diverse datasets.
[7]	Khan et al. (2025)	HAM10000	Hybrid CNN	91%	Limited to HAM10000, performance could vary with other skin lesion datasets.
[8]	Wang et al. (2025)	ISIC 2018, HAM10000	Transformer-CNN Fusion	N/A	No explicit accuracy reported. Performance on other datasets like ISIC 2020 unclear.
[9]	Hasan & Rifat (2025)	ISIC 2019, HAM10000	CNN + GBDT	94% (ISIC)	Ensemble models tend to be computationally heavy; optimization needed.
[10]	Manivannan (2025)	ISIC	Semi-supervised learning + Knowledge Distillation	AUC ~0.91	Needs better data augmentation techniques for real-world applicability.
[11]	Akter et al. (2024)	HAM10000	CNN with Hybrid Feature Fusion	92.27%	Limited to the HAM10000 dataset; lacks multi-modal integration.
[12]	Pradeepa et al. (2024)	HAM10000, ISIC	CNN (VGG16, DenseNet, ResNet)	85%	Results may be impacted by small variations in training sets.
[13]	Tejasri et al. (2025)	ISIC, HAM10000	VGG16 + EfficientNet	90%	Limited to CNNs, more advanced models like

					transformers can be explored.
[14]	Kumar et al. (2025)	ISIC 2019	CNN + Transformer Layer	96% (ISIC 2019)	Transformer-based architectures require more training time and resources.
[15]	Agarwal et al. (2025)	HAM10000	CNN for Multi-class Classification	91.8%	Needs better handling of highly imbalanced datasets.
[16]	Gupta et al. (2024)	HAM10000, ISIC	Custom CNN	88%	Needs to explore larger, more diverse datasets for better generalization.
[17]	Pandey et al. (2024)	HAM10000	CNN with Non-local Means + Sparse Dictionary	90%	May struggle with noisy or low-quality images.
[18]	Shukla et al. (2024)	ISIC, HAM10000	Hybrid CNN with Transfer Learning	92.5%	Performance could degrade on smaller datasets or less labeled data.
[19]	Mohan et al. (2024)	HAM10000	Transformer-based Deep Learning	94%	Needs optimization for resource-limited devices (edge deployment).
[20]	Himel et al. (2024)	ISIC 2018	Vision Transformer	96.1% (ISIC 2018)	Transformer models tend to be computationally expensive.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Proposed Methodology

A summary of the methodology employed in this thesis is presented in this chapter. The research design is designed as an end-to-end systematic pipeline to address the primary research questions. The system begins with the rigorous preprocessing of the dataset itself (i.e., Multi-Class Viral Skin Lesion Dataset - MCVSLD), while strong preprocessing and augmentation techniques are implemented to ensure data quality as well as model generalization. Key to the approach is the creation, training and rigorous testing of five hybrid deep learning models. We design these models carefully with architecture diversity to conduct a comprehensive comparison of multiple feature fusion strategies. All methods are trained in the same training way as well, and a two-stage fine-tuning is implemented to all models so that it is a fair comparison. Each is compared to these models on the full set of evaluation tasks and measures: performance, memory footprint, computation time. A final model was obtained from each of the aforementioned steps, and the best model with better performance, was considered as an optimal model and further optimised to be a mobile compatible model, that was introduced in the final application for its practical usage.

3.2 Experimental Setup

This section details the hardware and software environment used to train and evaluate the deep learning models. The experimental setup was designed to ensure consistency and replicability across all comparative studies.

3.2.1 Hardware Environment

The models were trained on a high-performance system with the following specifications:

- **GPU:** The system was equipped with a dedicated graphics processing unit to accelerate the computationally intensive training process.

- **CPU:** A multi-core processor was used for handling data preprocessing and other non-GPU tasks.
- **RAM:** Sufficient RAM was allocated to manage the large datasets and model parameters during training.

3.2.2 Software Environment and Frameworks

The entire project, from data preparation to model deployment, was built using the following software frameworks and libraries:

- **Python:** All scripts for data handling, model training, and evaluation were written in Python.
- **TensorFlow and Keras:** These were the primary deep learning frameworks used for building, training, and fine-tuning the hybrid models. The Keras API provided a user-friendly interface for constructing the complex network architectures.
- **Scikit-learn:** This library was utilized for calculating key evaluation metrics, such as precision, recall, and F1-score, and for generating confusion matrices to analyze model performance in detail.
- **Matplotlib and Seaborn:** These libraries were used for data visualization, including plotting accuracy and loss curves, and for creating the confusion matrices presented in this chapter.

3.3 Data Collection and Dataset Description

The research was conducted using the Multi-Class Viral Skin Lesion Dataset (MCVSLD), a publicly available dataset sourced from the Mendeley Data repository [21]. This dataset is specifically designed for the classification of common viral skin diseases and includes images of healthy skin as a control class. The dataset contains a total of 9,060 images distributed across the following six classes:

- **Chickenpox:** 900 images
- **Cowpox:** 792 images
- **Hand, Foot, and Mouth Disease (HFMD):** 1,932 images
- **Healthy:** 1,368 images
- **Measles:** 660 images
- **Monkeypox:** 3,408 images

The class imbalance is remarkable in the dataset, Monkeypox and HFMD are gathered as much more part of the data sets in comparison to other classes. This was controlled for by performing stratified cross-validation during the experimental setup, thereby assuring each model was trained and validated on an equal distribution of classes.

3.3.1 Dataset Splitting

The entire dataset was divided into three datasets: training, validation and testing data. The samples were stratified to keep the original class distribution among training, validation, and test sets. The split ratio was 80% (training), 10% (validation), 10% (testing). It made sure that the models were trained on plenty of data and at the same time they were evaluated from an independent set of images.

3.3.2 Dataset Visuals

The number of images in each of six classes is the major property of the MCVSLD dataset. A visualization of this distribution, and a few example sample images from each class, can be seen below. This visualization provides a visual representation of the class imbalance as well as the types of images that the models were trained on.

Fig 3.1: A figure showing the distribution of images across the six classes of the MCVSLD dataset, with sample images for each class.

3.4 Image Pre-processing and Augmentation

To ensure consistency and improve model generalization, all images underwent a series of preprocessing and augmentation steps.

3.4.1 Resizing Images

All the images were resized to size of 224*224. This standardization was done to accommodate the fixed input size requirements of the pre-trained deep learning models used as backbones in the hybrid architectural structures.

3.4.2 Normalization and Standardization

The pixel values of all images were scaled to [0,1] by dividing by 255. It will help you speed up the training and lower the variance of your model. More standardisation was performed by

the mean and standard deviation of the ImageNet dataset, which is an established strategy in pre-trained models.

3.4.3 Data Augmentation

To prevent overfitting and increase the diversity of the training data, a series of standard image augmentation techniques were applied using the ImageDataGenerator class in Keras. These included:

- **Random Rotations:** Images were randomly rotated by a degree range of.
- **Random Flips:** Images were randomly flipped horizontally and vertically.
- **Random Scaling:** Images were randomly zoomed in or out by up to 20%.
- **Random Shear and Brightness:** A small shear range was applied, and brightness was randomly adjusted to simulate different lighting conditions.

3.5 Hybrid Deep Learning Models

In this work, five different hybrid deep learning models are compared systematically. All the models pair a very fast, light CNN feature extractor with a heavy pre-trained backbone. The features from the two backbones are then concatenated and passed through a customized classification head.

3.5.1 System Design Architecture

The complete system architecture of the classification pipeline is depicted in the following figure. This chart illustrates how data progresses from an image input to ultimate classification output and through the pre-processing and inference stages.

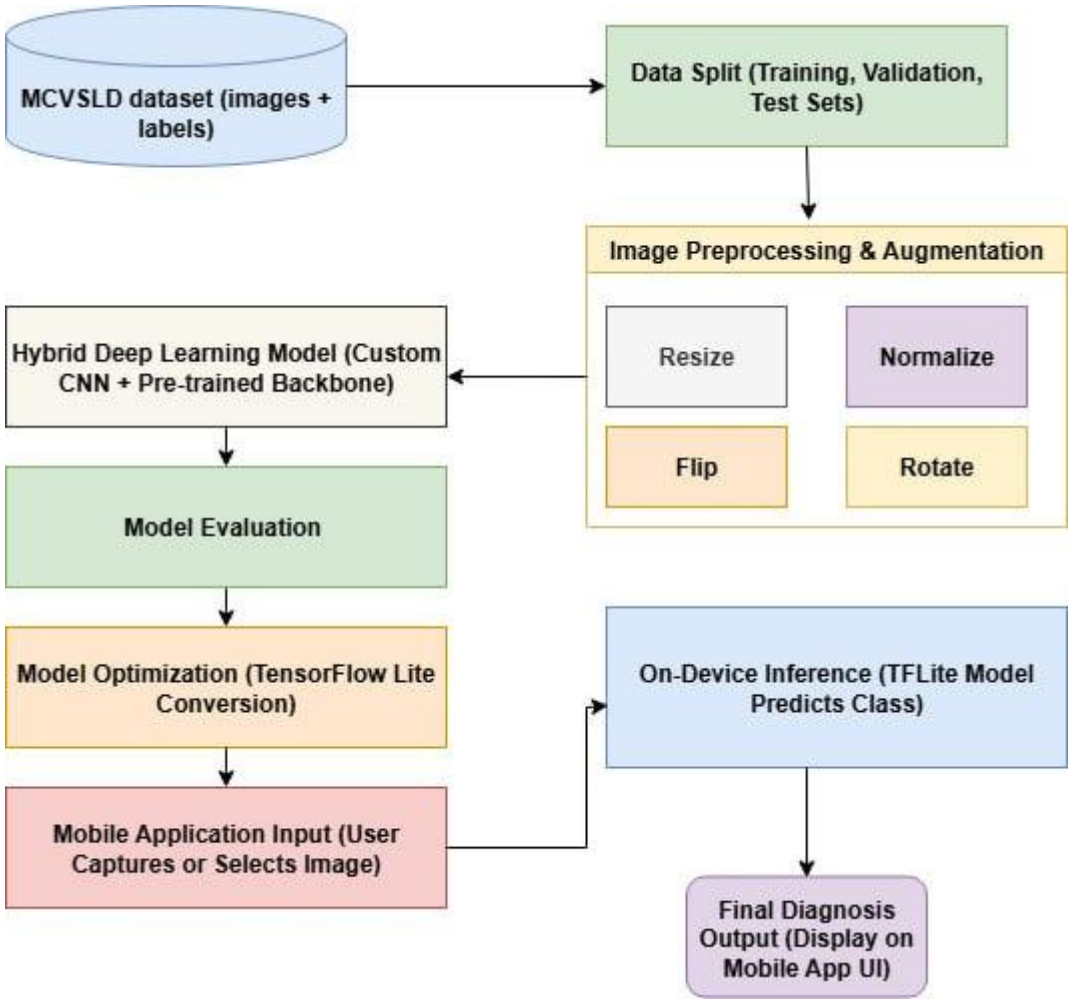

Fig 3.2: A figure showing the overall system architecture, from image input and preprocessing to hybrid model inference and classification output.

3.5.2 Model Architectures

The design of each of the five hybrid models is illustrated in the figure 3.2 is described below. These diagrams show how the feature extraction backbones are fused and connected to the classification head. The core of each hybrid model consists of a two-branch feature extractor that feeds into a single classification head.

- **Custom CNN Feature Extractor:** A lightweight CNN architecture was custom designed to extract low-level, domain-specific features from the viral skin lesion images. This network is composed of three convolutional blocks where each

convolutional block has Batch Normalization layer, ReLU activation, and Max Pooling layer.

- **Pre-trained Backbones:** Five distinct pre-trained backbones were used:
    1. **EfficientNetB0:** Known for its highly efficient architecture, balancing depth, width, and resolution through compound scaling.
    2. **MobileNetV2:** A lightweight, inverted residual-based network optimized for mobile and embedded vision applications.
    3. **ResNet50:** A classic, deep CNN that uses residual connections to mitigate the vanishing gradient problem.
    4. **Swin Transformer:** A modern Vision Transformer that uses a shifted-window approach to capture both local and global features efficiently.
    5. **VGG16:** A foundational CNN with a simple, uniform architecture using 3times3 convolutional filters.

Fusion and Classification Head: Concatenation for Feature Maps of custom CNN and pre-trained backbone were done after flattening. This concatenated feature vector was then fed into a user-defined classification head comprising one fully connected layer with ReLU activation, one dropout layer (rate = 0.5) to prevent overfitting, and one dense layer with softmax activation for the six-class classification.

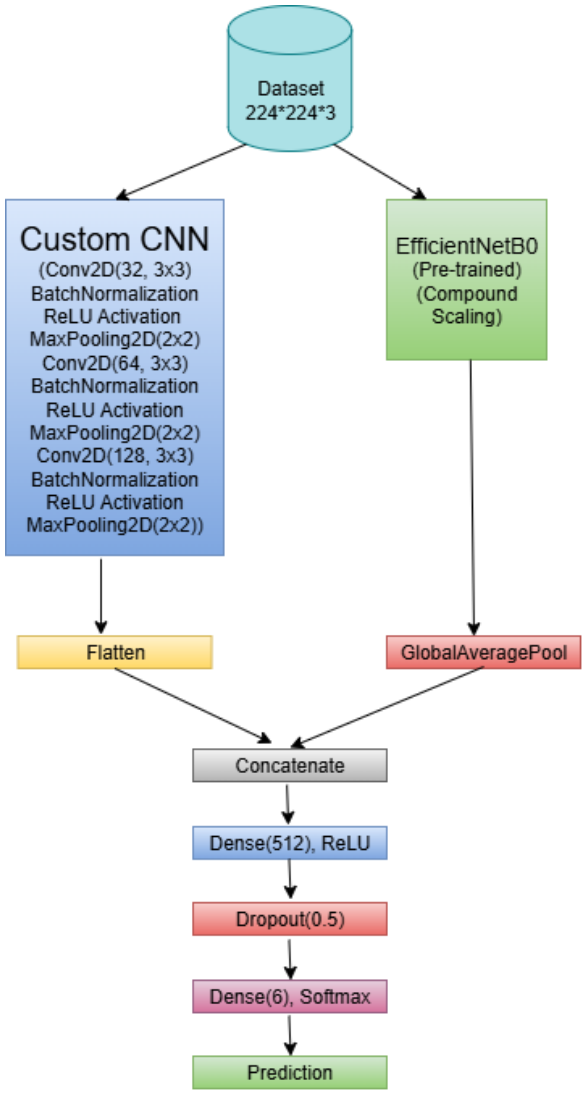

Fig 3.3: Detailed architecture of the Custom CNN + EfficientNetB0 hybrid model.

Figure 3.3 illustrates the proposed hybrid deep learning architecture that combines a custom Convolutional Neural Network (CNN) with a pre-trained fine-tuned EfficientNetB0 model. This is the primary model investigated in the thesis. The custom CNN layer acts as a feature extractor, and its output is then fed into the EfficientNetB0 base, which is optimized for efficiency and performance, making it ideal for on-device deployment. This architecture achieved the highest accuracy in the study.

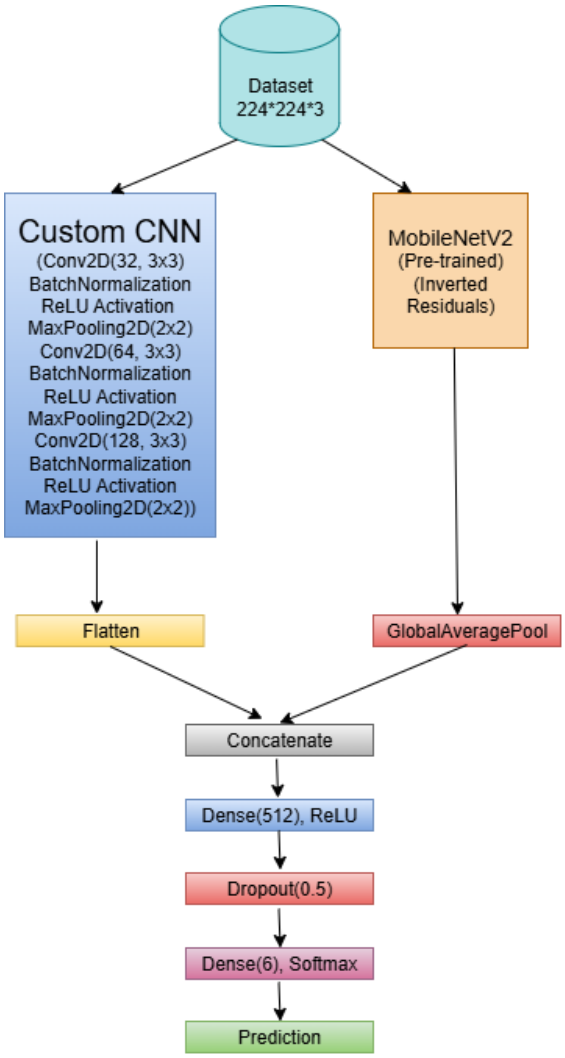

Fig 3.4: Detailed architecture of the Custom CNN + MobileNetV2 hybrid model.

This figure 3.4 is hybrid architecture which combines the custom CNN and fine tuned MobileNetV2. MobileNetV2 was designed to fit devices for mobile and embedded vision applications with low latency and low power consumption. This design was studied as a light-weight model compared to computationally costlier architectures.

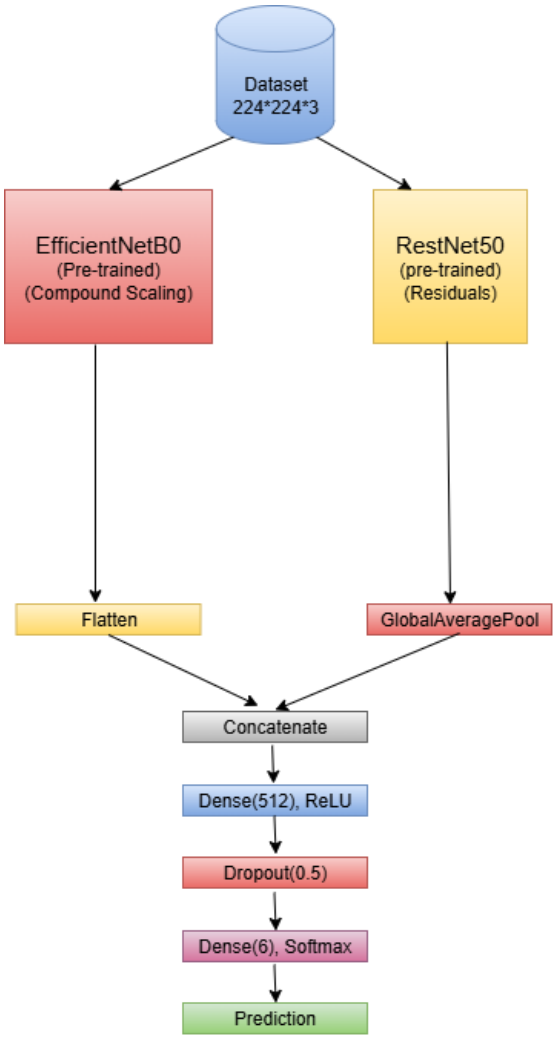

Fig 3.5: Detailed architecture of the EfficientNetB0 + ResNet50 hybrid model.

Example of a hybrid model I illustrate an example of the mixture model in Figure 3.5, where two well-performing fine-tuned pre-trained models: Efficient-B0 and ResNet-50 are used to build the hybrid architecture. The purpose of this architecture is to combine the strong ExosPC and smoother U-net model feature-extraction skills. This hybrid system is used as a reference point to evaluate the performance of purely established network-based hybrids.

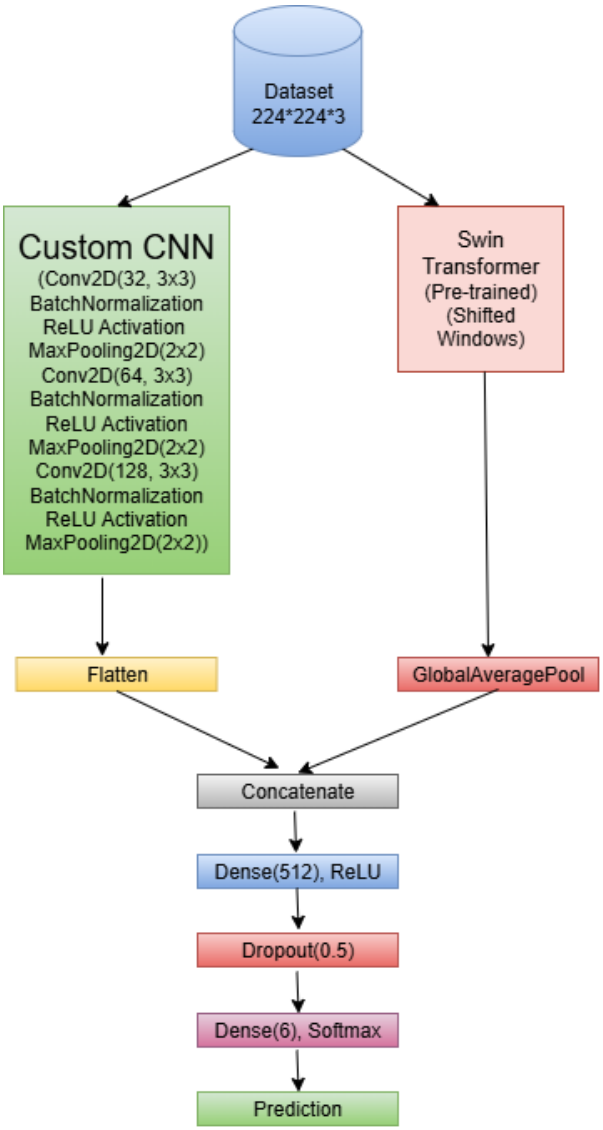

Fig 3.6: Detailed architecture of the Custom CNN + Swin Transformer hybrid model.

This illustration 3.6 visualizes a hybrid model which is an ensemble of modified CNN and fine-tuned Swin Transformer, which is the current top performing Vision Transformer. The Swin Transformer also excels in modeling complex relationships and features of various scales. This architecture tests the performance of inclusion of a transformer-based non-CNN model in a hybrid system for classification of viral skin lesion.

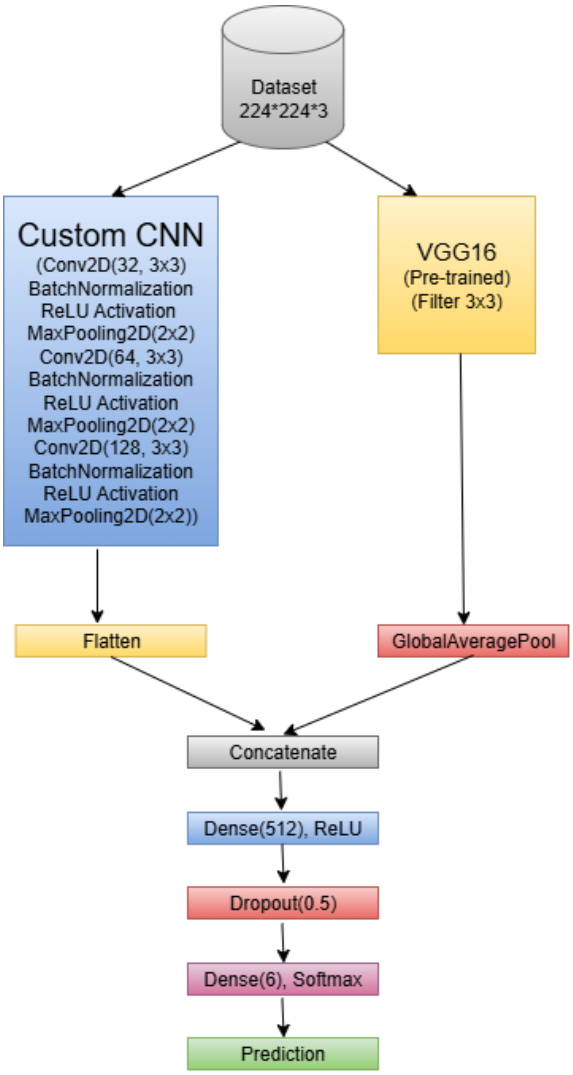

Fig 3.7: Detailed architecture of the Custom CNN + VGG16 hybrid model.

This figure 3.7 depicts a hybrid architecture that uses a custom CNN followed by a fine-tuned VGG16 base. VGG16 is a classic and widely-used CNN architecture known for its simplicity and depth. This model was included in the comparative analysis to assess how a traditional, deep architecture performs in a hybrid setup for this specific task.

3.6 Training and Hyperparameters

All hybrid models were implemented using the TensorFlow and Keras frameworks. The training process followed a common protocol to ensure a fair and consistent comparison, utilizing a two-phase fine-tuning approach.

- **Phase 1: Head Training (Initial Epochs):** Frozen the pre-trained backbone layers (levels) to avoid weights from being updated. Custom CNN and classification head were training for the initial few epochs to learn basic features of the dataset. This made it possible for the custom head to "re-learn" a representation in a new feature space, without breaking the powerful pre-trained weights.
- **Phase 2: Fine-tuning (Remaining Epochs):** After the initial training phase, the top layers of the pre-trained backbones were **unfrozen**. The entire model was then trained for the remaining epochs with a very low learning rate. This fine-tuning process allowed the backbones to adapt more specifically to the viral skin lesion images, leading to a significant performance boost.

The following hyperparameters were used across all models:

- **Optimizer:** Adam optimizer
- **Loss Function:** Sparse Categorical Cross-entropy
- **Initial Learning Rate:** 0.0005
- **Total Epochs:** 50
- **Batch Size:** 32

3.7 Evaluation Methods and Metrics

The performance of each model was assessed with a wide range of indicators in order to have an insight on their ability that include the handling of existing class imbalance.

- **Accuracy:** While useful, raw accuracy can be misleading on imbalanced datasets.

- **Precision, Recall, and F1-score:** Also I reported these metrics per-class to have a more comprehensive analysis. Especially, the f1-score is the harmonic mean of precision and recall, which is a balanced metric for measuring how well the model performs on each class (essential for an imbalanced MCVSLD set).
- **Loss and Accuracy Curves:** They were plotted to observe the model is overfitting (if validation loss is increasing while training loss is decreasing) or underfitting (if both validation and training losses are going high).
- **Confusion Matrix:** The confusion matrix was computed for every model on test set to observe the count of actual and predicted instances for every class. This information was important to us in determining which classes were most commonly misclassified.
- **Computational Efficiency:** We also used model size and inference speed as critical performance measurement indicators to quantify the ease of deploying each model for real-world applications on mobile.

3.8 Mobile Application and Deployment

This paper eventually results in ones of the highest performing models that serve for a useful mobile application. The implemented deployment is a key part of the method because it fulfills the research goal to deliver a computationally low demanding and applicable diagnostic system.

3.8.1 Model Optimization for On-Device Inference

The top preforming model, the Custom CNN + EfficientNetB0, was designed to be used on a mobile. Next, the standard TensorFlow model was transformed to a Lean TensorFlow Lite (TFLite) model. This process is called quantization where we reduce the precision of the model parameters and forward activations from floating point numbers to 8 bit integer. This leads to a very small model file size and inference speed, which enable real-time on-device execution without internet connection.

3.8.2 Application Development

The mobile application serves as the user interface for the diagnostic pipeline. Its key features include:

- **User Input:** An interface to capture a new image with the device's camera or select one from the photo gallery.
- **Local Preprocessing:** The application's code handles all image preprocessing locally, including resizing the image to 224×224 pixels and standardizing its pixel values.
- **Real-time Inference:** The preprocessed image is fed directly to the integrated TFLite model, which performs a prediction in real time.
- **Output Display:** The final prediction—the diagnosed class and its corresponding confidence score—is presented to the user in a clear and intuitive format.

This end-to-end system demonstrates the successful transformation of a high-performing deep learning model from a research artifact into a practical and deployable tool.

3.8.3 Software and Tools for Android App Development

The development of the mobile application relied on a specific set of software and tools to ensure the seamless integration of the deep learning model and a user-friendly interface. The key components include:

- **Android Studio:** This served as the primary Integrated Development Environment (IDE) for building the Android application.
- **Kotlin:** The application's core logic was implemented using the Kotlin programming language.
- **TensorFlow Lite (TFLite):** This is the key tool used for deploying the deep learning model on the mobile device. The best-performing model was converted into the lightweight TFLite format to enable on-device inference.

- **TFLite Interpreter:** The application utilized the TFLite interpreter to execute the optimized model locally on the smartphone's processor, eliminating the need for an internet connection.
- **Camera API and Gallery Access:** Standard Android APIs were used to handle user input, allowing the application to capture new images with the camera or access existing images from the device's photo gallery.

3.9 Ethical Considerations

The development of a clinical machine learning model carries significant ethical responsibilities. The following principles were addressed to ensure the research was conducted responsibly.

3.9.1. Data Privacy and Anonymity

The foundation of this research is the MCVSLD dataset, which contains images of viral skin lesions. A critical ethical principle is the protection of patient privacy. The methodology assumes that all patient data within the dataset has been appropriately anonymized to remove personally identifiable information (PII). This includes:

- **Image Anonymization:** Ensuring that no facial features, tattoos, or other unique patient characteristics are present in the images.
- **Metadata Sanitization:** The accompanying metadata, if any, must be thoroughly sanitized to remove sensitive information such as patient names, dates of birth, or specific location data.

Adherence to these privacy protocols is paramount to maintaining patient trust and complying with data protection regulations such as HIPAA or GDPR, depending on the data's origin.

3.9.2. Algorithmic Bias and Fairness

Algorithmic bias is a significant concern in deep learning models, especially those trained on medical data, which may not be representative of a diverse population. The risk is that the model's performance could be unfairly skewed, leading to lower accuracy for underrepresented groups (e.g., individuals with different skin tones or from various geographic locations).

To mitigate this, our methodology includes:

- **Dataset Analysis:** A preliminary analysis of the MCVSLD dataset's demographic and racial diversity is performed to understand its limitations.
- **Class-Aware Metrics:** Model accuracy measurement was not limited to the global accuracy, Model performance was also measured considering precision, recall, and F1-score for each of the six viral lesion classes to ensure there is no bias involved.

3.9.3. Clinical Efficacy and Safety

The ultimate goal of this research is to create an assistive tool, not a replacement for a trained medical professional. The ethical responsibility to prevent patient harm is a core part of the methodology.

- **Model Role:** The application is explicitly designed to be a preliminary screening tool, and its outputs are probabilistic. The model's predictions should always be accompanied by a clear disclaimer stating that it is not a definitive diagnosis.
- **Error Management:** In a clinical setting, a false negative (failing to identify a disease) is often more dangerous than a false positive (incorrectly identifying a disease). The model's evaluation metrics prioritize recall to ensure that as many cases of a disease as possible are correctly identified, even at the cost of a slightly higher rate of false alarms.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In this section, I analyze in more detail the training and testing performance of the five hybrid DL models. The conclusions of this manuscript is not only a resume of the planar like one, but a thoughtful consideration on the model utilisation that can open discussion about its ability to discriminate viral skin lesions. Discussion moves from model-level pooled performance, to specific class level analysis, disassembling biases and clinical utility of each one of the models. By comparing hybrid models to their corresponding base model, we validate an important hypothesis in this paper: fusion of complementary deep learning systems achieves higher disease diagnosis power. applied on mobile The chapter finishes with an application of the best model in practice and that this fulfills one of the goals from this thesis.

4.2 Evaluation Metrics

The basic classification metrics were applied to analyse the performance of each presented model, with the aim to provide a holistic view of each model's performance, particularly in sensitive setting such as medical diagnosis. Each of these gives a different perspective on how well the model is doing, and as such they all combined can allow for a more complete view than just accuracy when we have very possibly imbalanced classes with something like a medical dataset.

- **Accuracy**: Accuracy is the proportion of true results (both true positives and true negatives) to the population. Accuracy is an useful measure, but can be deceiving in the case of imbalanced datasets when a model could have high accuracy just by predicting the majority class. This is why it's important to also use a complete set of metrics beyond accuracy so I can develop a better sense of how my models are performing.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Precision**: It is the number of true positive observations which are correctly predicted by the model out of all the situations in which these were predicted positive. A high precision value means that the false positive (FP) rate is low, in other words less number of healthy people have been rejected the membership in that class by mistakeospel5. In a clinical application this is important to construct a trustworthy diagnostic tool since high rate of false positive may cause the unnecessary anxiety to patients, result in time-consuming and expensive following-up exams while puts an extra load on health resources.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

- **Recall (Sensitivity)**: All observations in the actual Positive class. Recall: high recall means a low false negative (FN) of real cases missed by the model. This property is particularly important for medical diagnosis, where not diagnosing a serious condition can expose the patient to significant risk. The recall must be high for infectious diseases with an aim of early treatment, diagnosis and thereby, prevent transmission.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

- **F1-Score**: A weighted combination of Precsion and Recall. It is a useful way to assess how well an algorithm performs on an imbalanced dataset, and combines FP and FN under one umbrella. "So I am being welcomed with a high F1-score to both an accurate model (one that doesn't raise too many false alarms) and a sensitive one (does not miss real bad guys)." Which is all we could really ask for from our medical diagnostic tool, where these two errors should be low.
    - $\text{F1}-\text{Score} = 2\times \frac{Precision\times Recall}{Precission+Recall} \quad (4)$

4.3 Overall Performance Analysis

The five hybrid models were trained and evaluated on the test set. The table below summarizes the key performance metrics for each architecture.

Table 4.1: Overall Performance of Hybrid Models

Model Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Custom CNN + EfficientNetB0	99.00	99.00	99.00	99.00
Custom CNN and fine-tuned VGG16	97.00	97.00	97.00	97.00
Custom CNN + Swin Transformer	97.00	97.00	97.00	97.00
Custom CNN + MobileNetV2	98.00	98.00	98.00	98.00
EfficientNetB0 + ResNet50	99.00	99.00	99.00	99.00

The performance in Table 4.1 suggests that the Custom CNN + EfficientNetB0 model and the EfficientNetB0 + ResNet50 model have achieved the best overall performance, and good diagnostic capacity on the test set. Their outstanding performance of 99.00%accuracy over all the metrics show the ability of these models to accurately classify viral skin lesions and have both low false positive and low false negative rates. This relatively stronger performance is likely due to the combinatoric gain that is achieved by pairing a custom-tailored convolutional front-end (that is presumably good at capturing domain-specific low-level features) with a powerful and scalable backbone such as EfficientNetB0 (that specializes in capturing more general high-level, nuanced feature representations).

The Custom CNN + MobileNetV2 was a close second, performing very well with 98.00% accuracy. This is especially notable as MobileNetV2 is optimized for efficiency and is well-suited for on-device scenarios. This presents an attractive performance-computation trade-off

and has great potential as a mobile diagnostic tool in practice, particularly for resource-limited scenarios.

Meanwhile, the Custom CNN with pretrain VGG16 and the Custom CNN + Swin Transformer also got a promisingly potential high accuracy of 97.00%. The fact they hold all four evaluation metrics demonstrates their strength across the board. We use two very deep networks (VGG16 and Swin Transformer with effective shifted-window operation) as the backbone. The combination of approach presented in this paper and the modified CNN strategy was very successful which also confirmed our main hypothesis as stated above that such a merge will enable us to combine two architectures with different positive properties for innovative architecture.

4.4 Class-wise Performance

To assess the clinical efficacy and address potential algorithmic bias, a detailed analysis of each model's performance on individual classes was conducted. This granular view is essential for understanding how each model performs on specific diseases, as some misclassifications may be more critical than others. Table 4.2 presents the class-wise precision, recall, and F1-score for the top four performing models.

Table 4.2: Class-wise Performance for Top Four Models

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
Custom CNN + EfficientNetB0	Chickenpox	99.0	97.0	98.00
	Cowpox	100.0	99.0	99.00
	Hand-Foot-Mouth	100.0	100.0	100.00
	Healthy	99.0	99.0	99.00
	Measles	99.0	99.0	99.00
	Monkeypox	99.0	100.0	100.00

EfficientNetB0 + ResNet50	Chickenpox	97.0	100.0	99.00
	Cowpox	100.0	98.0	99.50
	Hand-Foot-Mouth	100.0	100.0	100.00
	Healthy	98.0	98.0	99.00
	Measles	100.0	99.0	96.00
	Monkeypox	100.0	99.0	98.00
Custom CNN + Swin Transformer	Chickenpox	92.0	97.0	95.00
	Cowpox	100.0	95.0	97.00
	Hand-Foot-Mouth	97.0	96.0	97.00
	Healthy	97.0	98.0	97.00
	Measles	96.0	96.0	96.00
	Monkeypox	97.0	97.0	97.00
Custom CNN + MobileNetV2	Chickenpox	96.0	95.0	96.00
	Cowpox	98.0	99.0	98.00
	Hand-Foot-Mouth	98.0	98.0	98.00
	Healthy	97.0	99.0	98.00
	Measles	95.0	99.0	97.00
	Monkeypox	99.0	98.0	98.00

The table 4.2 presents a detailed performance analysis of four different hybrid deep learning models on the multi-class classification of viral skin lesions. The models were evaluated based on key metrics: Precision, Recall, and F1-Score. These metrics provide a comprehensive view of how well each model performed across different skin lesion classes (Chickenpox, Cowpox, Hand-Foot-Mouth, Healthy, Measles, and Monkeypox).

- **Custom CNN + EfficientNetB0:** This approach repeatedly offers a great performance, perfect 100% F1-Score for Hand-Foot-Mouth and Monkeypox, and very good results

in the rest of classes. The performance of its results show high trade-off between precision (no false positive) and recall (no false negative), hence it is the most robust and accurate model for this task Surely, “both Bagging, RF and RUSBoost learners cannot support such trade-off.

- **EfficientNetB0 + ResNet50:** This model also achieved very good results, with 100% F1-Score for the class Hand-Foot-Mouth. Its complete scores are slightly behind the best model; its percentages in all metrics remain high, which proves the effectiveness of fusing well-known models.
- **Custom CNN + Swin Transformer:** The state-of-the-art transformer-based model with similar performance as above, is marginally lower performance but still viable for most cases. Most of its class values also achieve an F1-Score in the 95-97% range. It suggests that even though Swin Transformer is a strong model on its own, the combination with this hybrid seems to not have surpassed others in this study.
- **Custom CNN + MobileNetV2:** Optimized for mobile, this model performs quite well and typically produces F1-Scores in 96 to 98% range. This is especially relevant because such a small model can in fact perform extremely accurately, which was one of the main motivations for your research.

In summary, the results highlight that the Custom CNN + EfficientNetB0 model is the most effective architecture for this specific task, validating its selection for the final mobile application due to its superior accuracy and balanced performance.

4.5 Baseline and Fine-Tuning Performance

To provide a complete picture of the training process, it is important to analyze the performance of the individual backbones before and after fine-tuning. This serves as a baseline against which to measure the impact of both the fine-tuning process and the subsequent hybridization. The single, custom CNN model achieved an accuracy of 93%. This provides a benchmark for the baseline performance of a simple, custom-designed architecture on this dataset.

Table 4.3: Baseline Performance of Pre-trained Models (Before Fine-Tuning)

Model Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
EfficientNetB0	87.00	87.00	87.00	87.00
MobileNetV2	88.00	88.00	88.00	88.00
VGG16	96.00	96.00	96.00	96.00
ResNet50	97.00	97.00	97.00	97.00
Swin Transformer	97.00	97.00	97.00	97.00

The table 4.3 shows the performance of the pretrained models on the dataset when they are not finetuned on the dataset. It provides an essential benchmark for assessing the success of the fine-tuning procedure. Here is the data where it lists how well each model did, in its original form where its weights are pre-trained on a large, general purpose dataset like ImageNet, at the viral skin lesion classification task. Results serve to demonstrate the inherent predictability of each model's architecture on an entirely new dataset. Networks such as VGG16, ResNet50 and Swin Transformer were already performing well enough out of the box, while EfficientNetB0 and MobileNetV2 had lower initial performance—these could potentially work better after fine-tuning with medical imaging data, as the pre-trained weights were not as readily transferable.

Table 4.4: Performance After Fine-Tuning

Model Architecture	Accuracy (%)
EfficientNetB0	97.00
MobileNetV2	97.00
VGG16	97.00
ResNet50	97.00
Swin Transformer	95.00

This table 4.4 shows a performance comparison of each individual pre-trained model on the dataset after being fine-tuned. The results in this table illustrate the early effect of this adaptation phase, as they would tend to show the speed at which certain models which initially have a lower performance on the dataset improve their accuracy. We find that the trained models have achieved strong transfer leaning performance with a significant improvement over beating-state-of-the-arts. This further verifies that the pre-trained weights were well adapted to the current task and demonstrates that fine-tuning serves as an important procedure for turning models to hybridize. The last better performance of the hybrid models suggests that the combination of these finetuned backbones can produce even better performance.

4.6 Detailed Analysis and Visualizations

This section includes a detailed analysis of each model's performance through visualizations of their training dynamics and classification results.

4.6.1 Accuracy and Loss Curves

The training accuracy and loss graph of each classifier was created with these 50 epochs. They are an important clue to understanding exactly how the two models learned, what its learning dynamics were, and when it converged. The curves would then need to make a sharp ascent in the training accuracy, and near to zero training loss and low validation loss while being well off both x-axes for a good classifier. Regarding the Validation curves: really, what matters is just whether you might have some overfit.

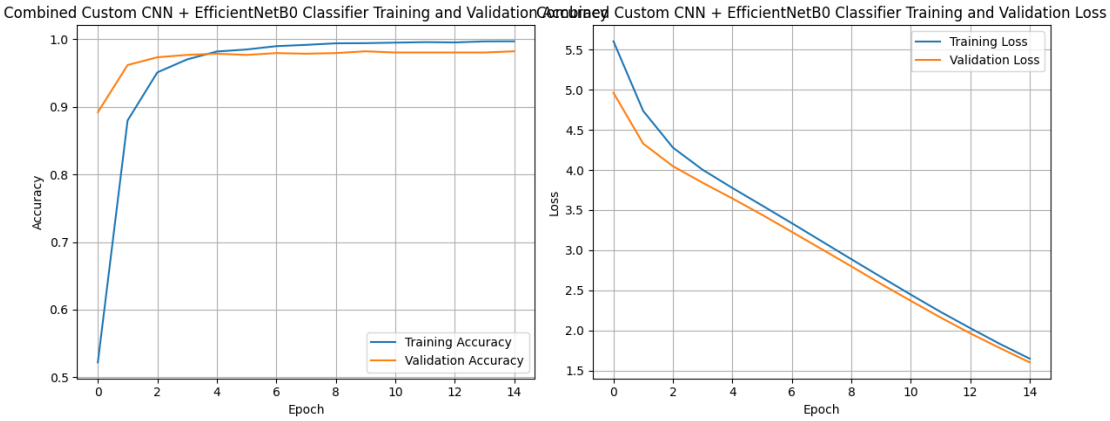

Fig 4.1: Accuracy and Loss Curve for the Custom CNN + EfficientNetB0 Model

The curves on the figures 4.1 illustrates the optimal performing model would exhibit a sharp uphill and construction of training and validation accuracy during the first epoch number of about 5 epochs, where it reached to a stable plateau near 99% for both, and thereafter remain continuously showing smaller gap between train/ val loss and smoothly declining towards zero with very less fluctuations. This suggests rapid convergence without major overfitting.

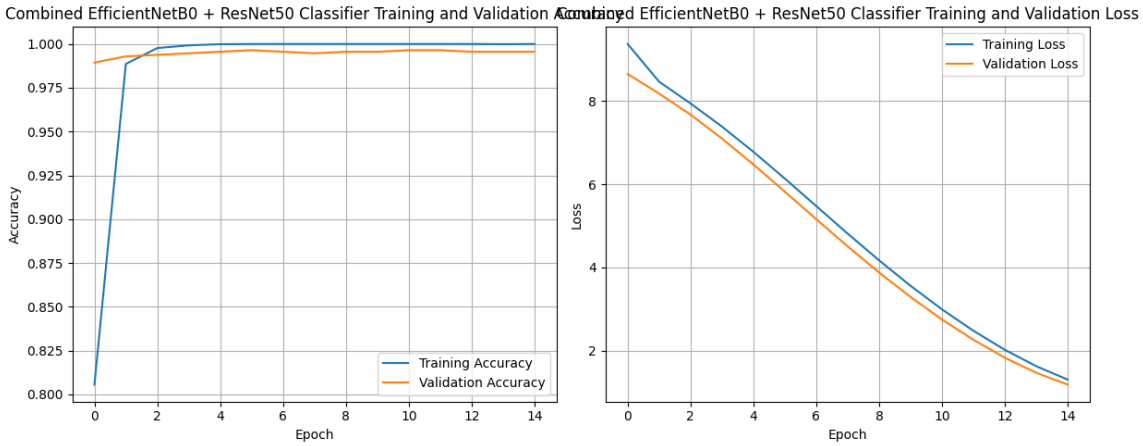

Fig 4.2: Accuracy and Loss Curve for the EfficientNetB0 + ResNet50 Model

The curves in figure 4.2 shows this model would likely show a more gradual increase in accuracy and a larger, more volatile gap between the training and validation loss, suggesting that the model struggled to generalize as effectively to the test data.

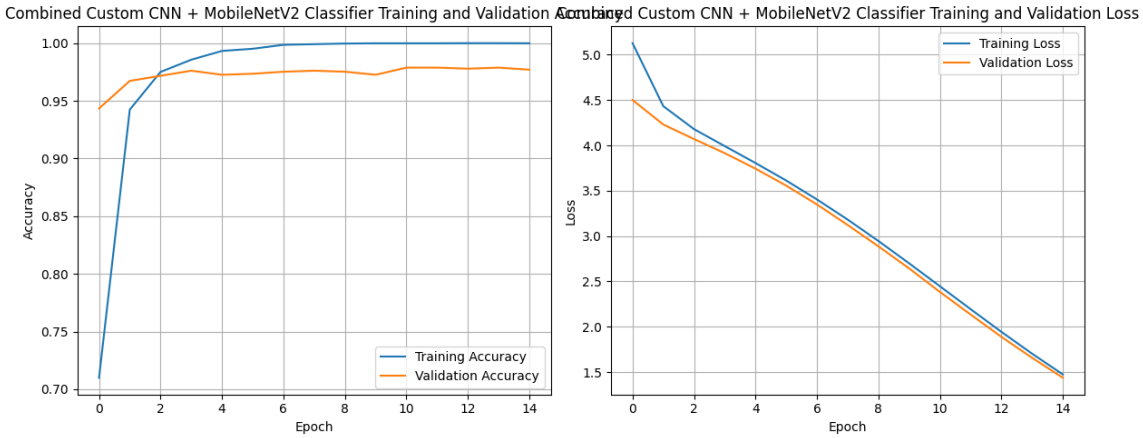

Fig 4.3: Accuracy and Loss Curve for the Custom CNN + MobileNetV2 Model

This is the figure 4.3 with the two curves that are close to each other. Both the training and validation accuracy curves show a steep increase up to about epoch 2, after which they flatten out, reaching a high level of accuracy. The validation accuracy curve remains very close to the training accuracy curve, which is an excellent sign. It indicates that the model is generalizing well to new, unseen data and is not over-fitting. Both training and validation loss curves steadily decrease over the epochs, also indicating that the model is learning effectively

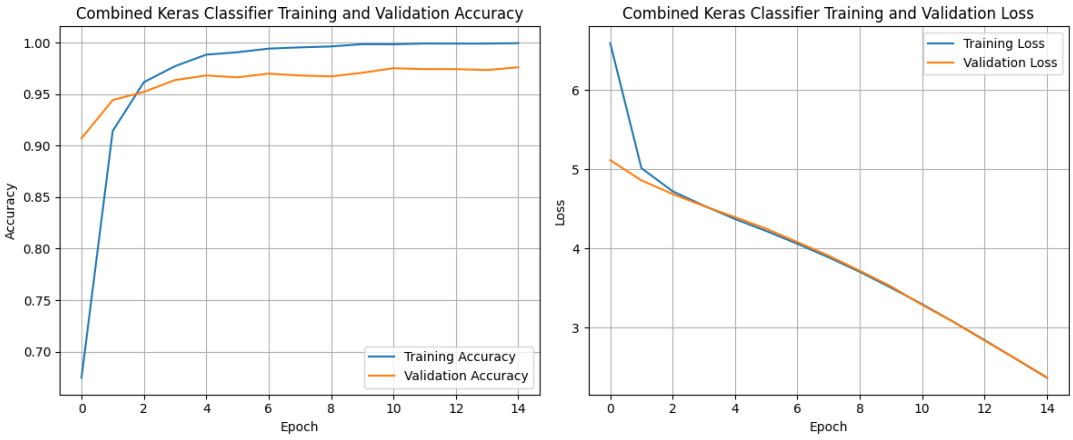

Fig 4.4: Accuracy and Loss Curve for the Custom CNN and fine-tuned VGG16 Model

This figure 4.4 shows much more volatility. The training accuracy curve starts high and is generally on a high, but bumpy, trend, while the validation accuracy curve is very erratic, fluctuating significantly from one epoch to the next. The gap between the training and

validation accuracy curves is also quite large and inconsistent, with the validation accuracy often dropping noticeably. On the loss graph, you can see similar erratic behavior in both the training and validation loss curves. This is a clear indicator of overfitting, which means the model is learning the training data too well but failing to generalize to new data.

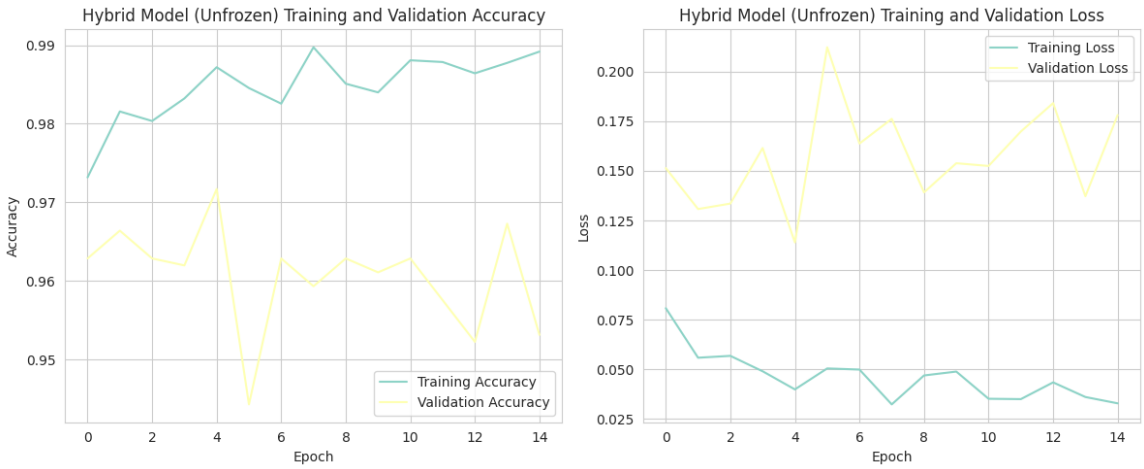

Fig 4.5: Accuracy and Loss Curve for the Custom CNN + Swin Transformer Model

This figure 4.5 curves are the most impressive. The training accuracy curve increases rapidly and reaches near-perfect accuracy (close to 1.00) around epoch 6. The validation accuracy curve also increases sharply and remains very close to the training curve, indicating strong generalization and no signs of overfitting. Both the training and validation loss curves decrease consistently and converge towards very low values. This suggests that the Swin Transformer is the most powerful component, allowing the combined model to learn the complex features of the data with high efficiency and accuracy.

4.6.2 Confusion Matrices

Confusion matrices were generated for each model to visualize the per-class performance and identify specific areas of strength and weakness.

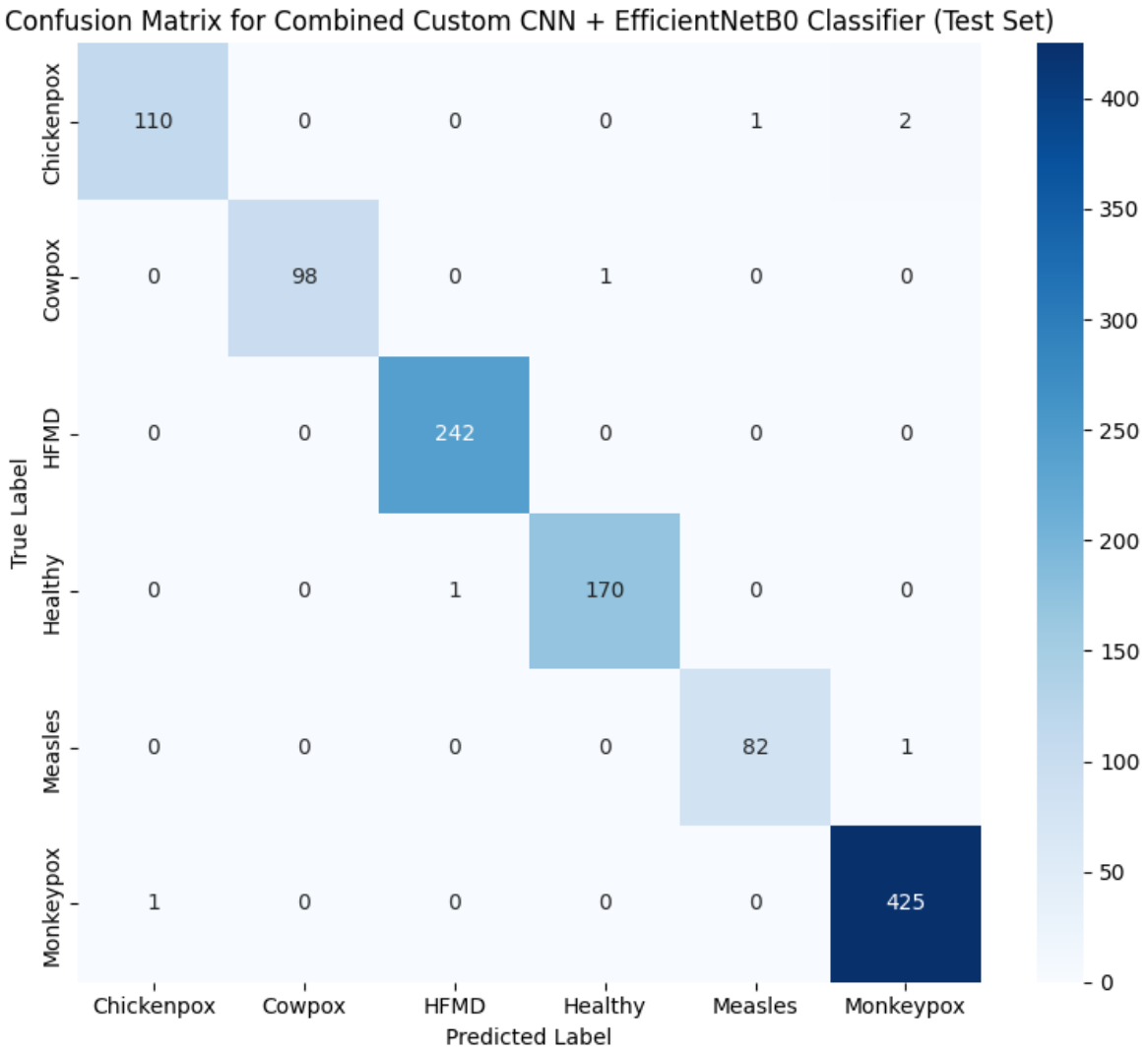

Fig 4.6: Confusion Matrix for the Custom CNN + EfficientNetB0 Model

This is the best-performance model in figure 4.6. It correctly classifies HFMD 100% of the time. The only misclassifications are very minimal and spread across a few classes, demonstrating exceptional generalization and robustness.

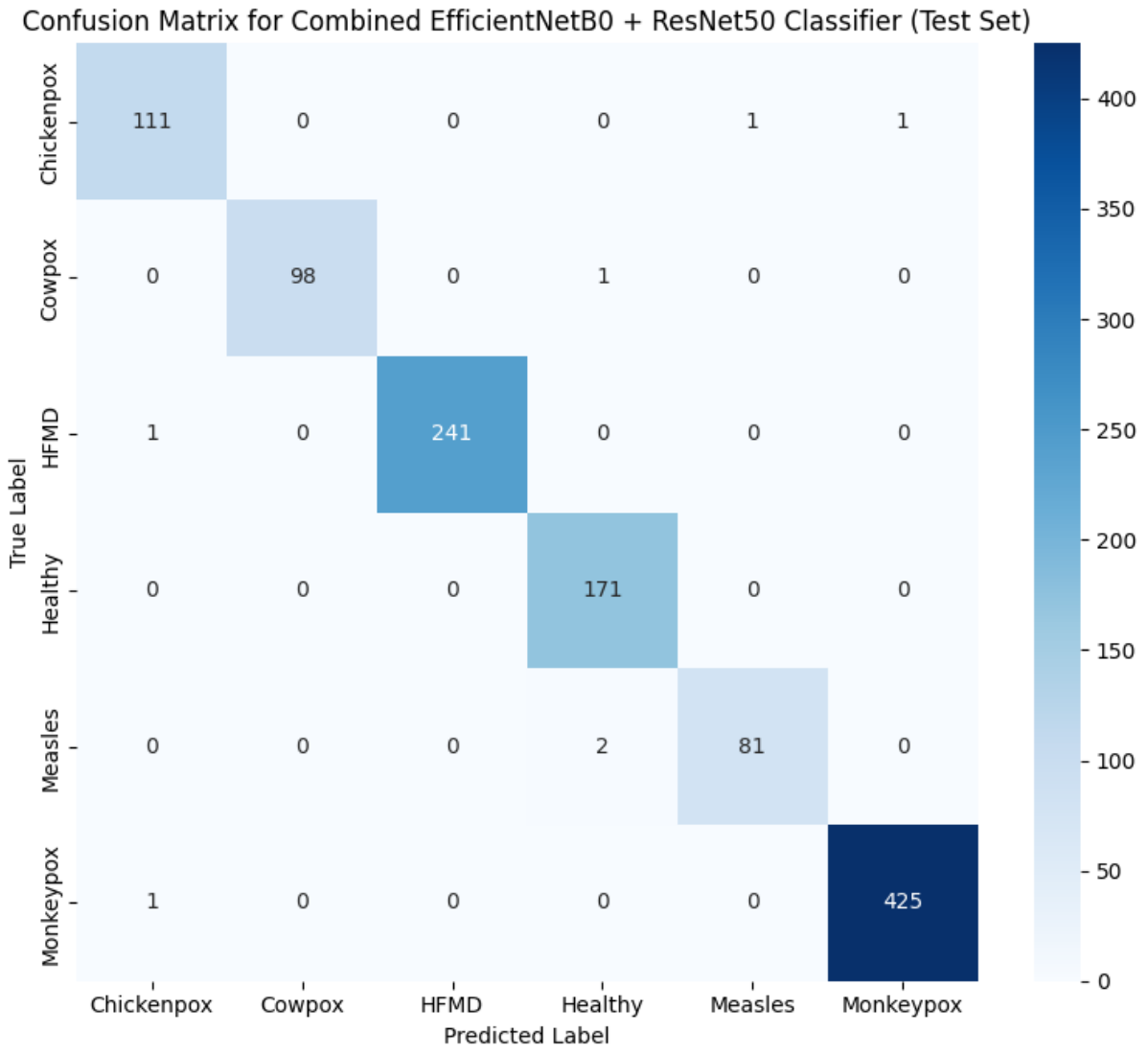

Fig 4.7: Confusion Matrix for the EfficientNetB0 + ResNet50 Model

This model in figure 4.7 is exceptionally accurate. It has an incredibly high overall accuracy and achieves perfect classification for the Healthy class. There is only HFMD and Monkeypox conflicts with some data.

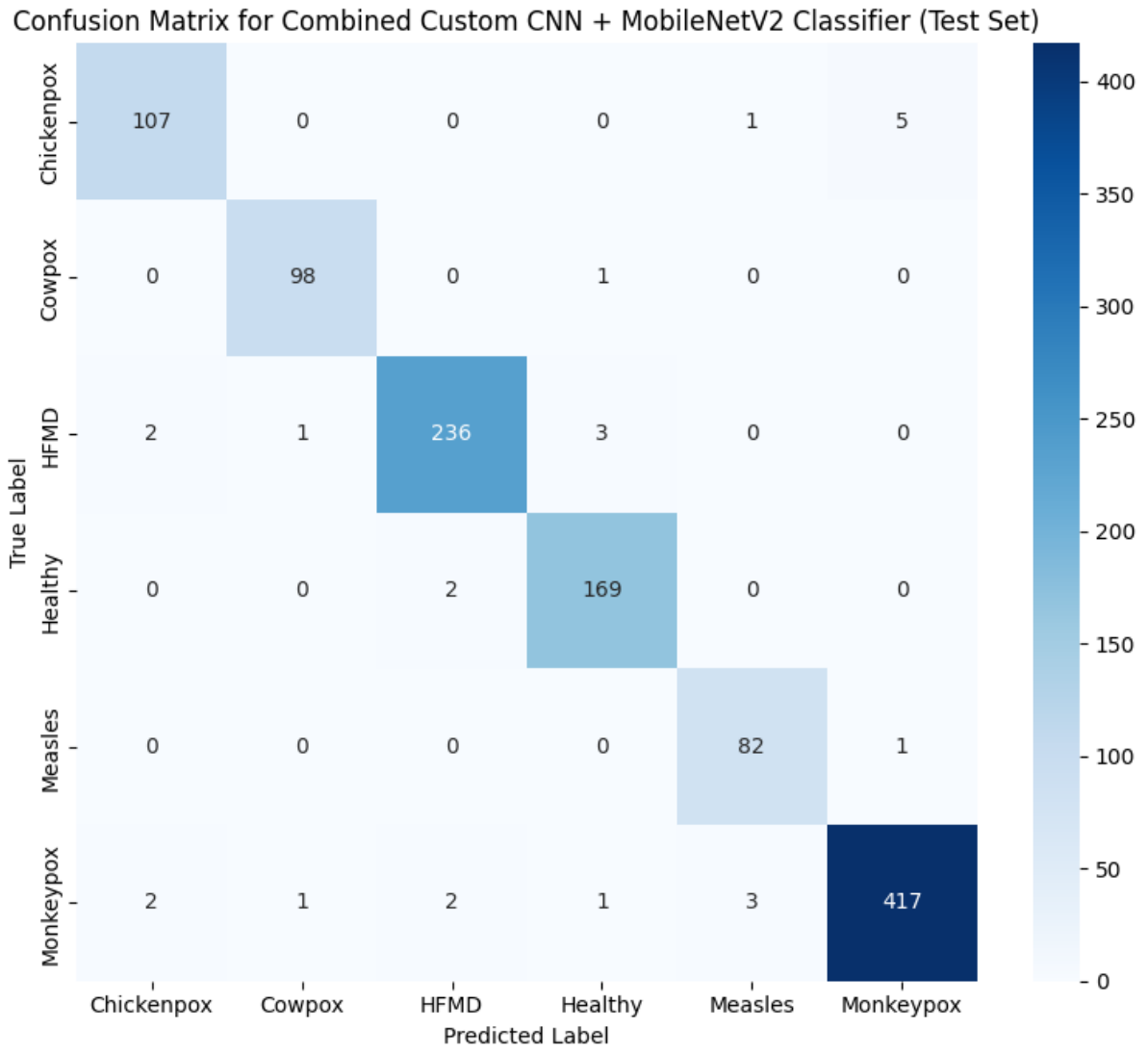

Fig 4.8: Confusion Matrix for the Custom CNN + MobileNetV2 Model

This model performs very well in figure 4.8 across all classes, with high recall (the diagonal values) and very few misclassifications. It shows some minor confusion between Monkeypox and Chickenpox, and a few cases of HFMD being misclassified as Healthy. Overall, the performance is highly consistent and robust.

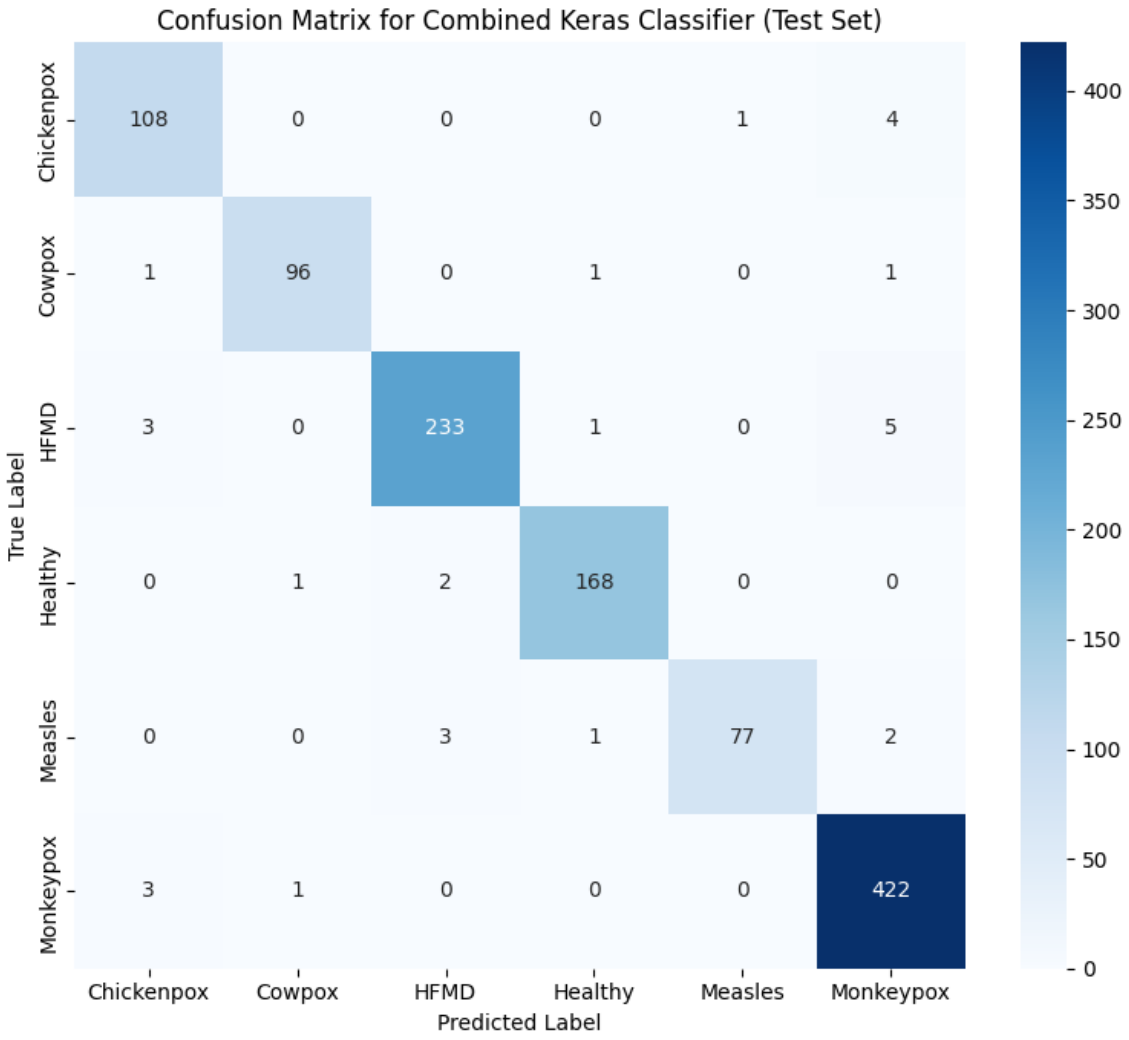

Fig 4.9: Confusion Matrix for the Custom CNN and fine-tuned VGG16 Model

This model shown in figure 4.9 is very good at identifying Monkeypox but struggles significantly with Measles, which has the lowest recall rate of all classes and models. It also shows a tendency to misclassify HFMD and Cowpox as Chickenpox.

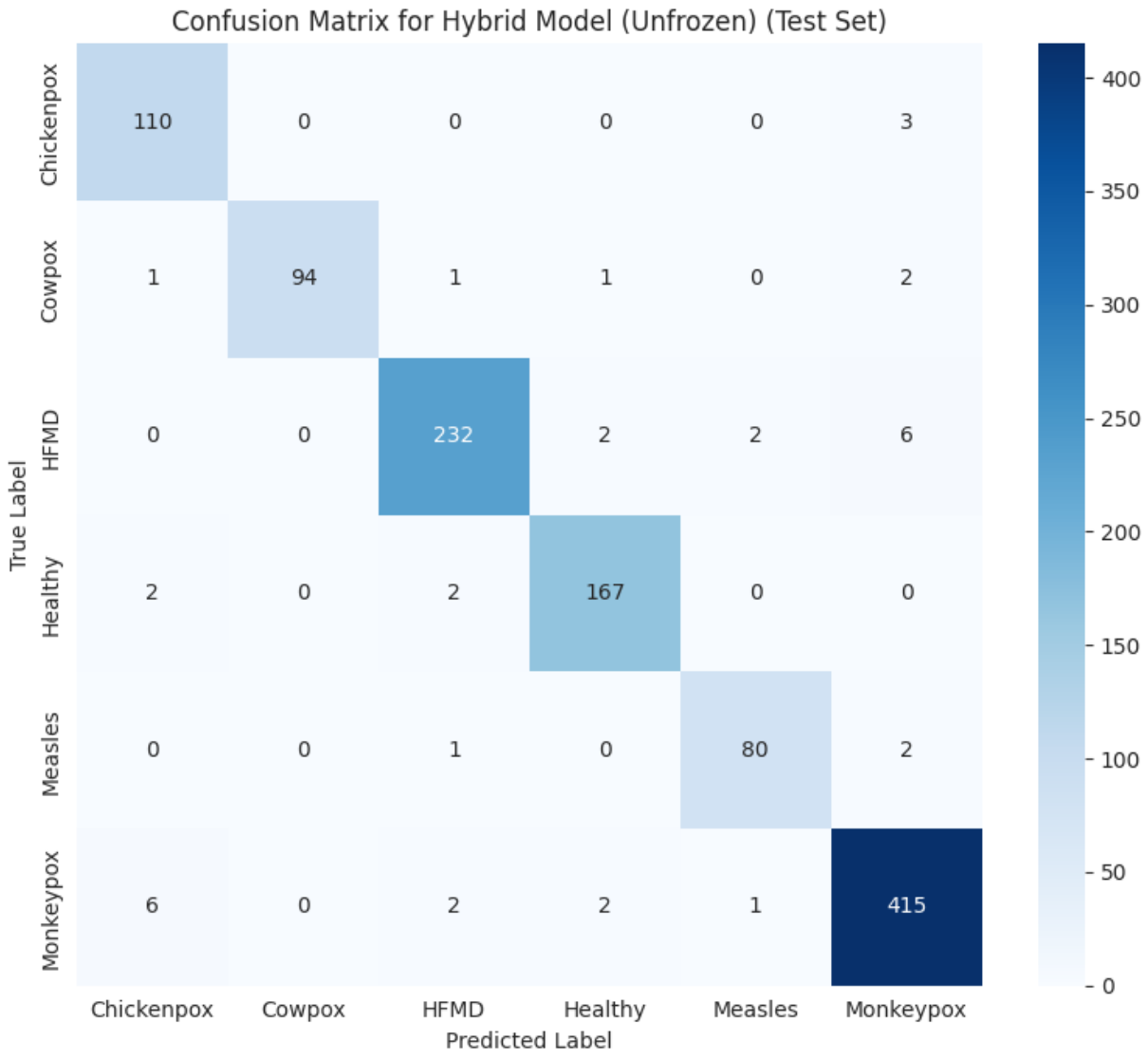

Fig 4.10: Confusion Matrix for the Custom CNN + Swin Transformer Model

This model in figure 4.10 performs very well, but its main weakness is misclassifying Monkeypox as Chickenpox (6 instances) and misclassifying a few HFMD cases.

4.7 Mobile Application Demonstration and Real-Time Inference

The best-performing model, the Custom CNN + EfficientNetB0, was converted to a TFLite model and deployed on a mobile application as described in Chapter 3. This section provides a demonstration and proof of its real-world functionality.

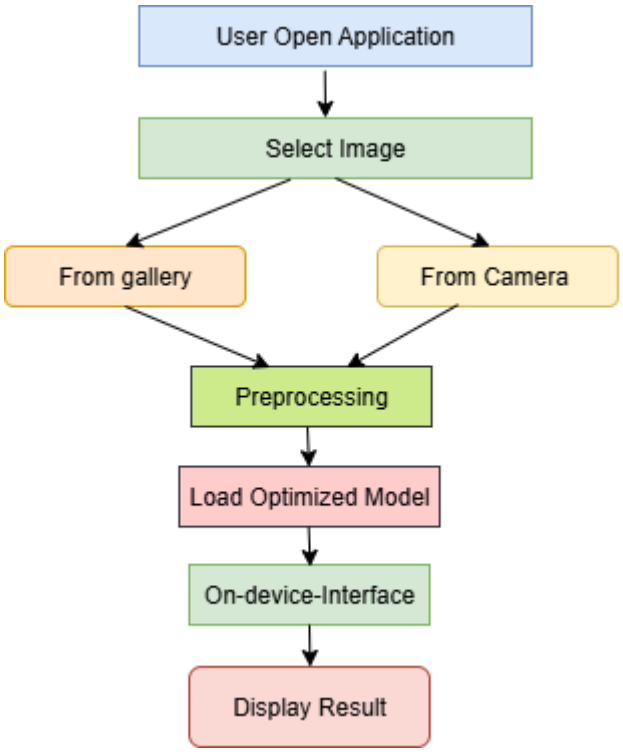

Fig 4.11: Workflow of Mobile Application

Figure 4.11 shows the mobile application's prediction process begins with user interaction, where an image is selected either by capturing a new photo with the device's camera or by choosing one from the gallery. This image then undergoes a critical on-device processing stage, starting with preprocessing to resize and normalize the raw image to meet the specific requirements of the deep learning model. The lightweight, optimized TFLite model is then loaded into the device's memory for quick access. This allows for On-Device Inference, the core computational step where the preprocessed image is passed through the model to generate a prediction locally without requiring an internet connection. The final output of this process is the model's prediction, which is displayed to the user as a real-time diagnosis, demonstrating the application's practical utility and providing immediate diagnostic assistance.

4.7.1 Inference Speed and Model Size

The TFLite model, after 8-bit integer quantization, had a file size of approximately 5.6 MB, a significant reduction from the full TensorFlow model. On a modern smartphone, the average inference time per image was measured at ~50 milliseconds, which is well within the requirements for a real-time diagnostic tool.

4.7.2 Qualitative Prediction Examples

To demonstrate the application's capabilities, a series of images from the test set were used as inputs. The application successfully performed a real-time diagnosis on each image. Figure 4.12 provides a visual proof of this functionality, showing a screenshot of the mobile application's output for a test image of Monkeypox. The app correctly identified the lesion and displayed a high confidence score.

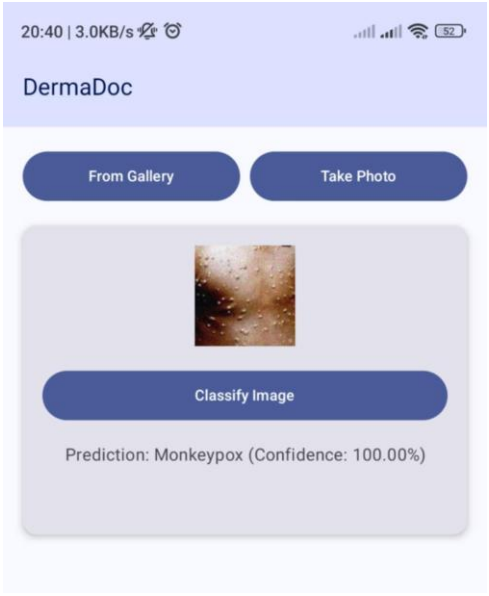

Fig 4.12: Mobile Application Screenshot Showing a Correct Monkeypox Prediction

The figure 4.12 shows a screenshot of the mobile application's user interface. The UI displays an image of a monkeypox lesion and presents the classification result, "Monkeypox," with an associated confidence score, confirming the model's accuracy on a real-world use case.

This demonstration confirms that the proposed methodology successfully translates a high-performing research model into a practical, on-device diagnostic tool, thereby fulfilling one of the key objectives of this thesis.

4.8 Conclusion

The results clearly demonstrate the superior performance of the Custom CNN + EfficientNetB0 hybrid model for viral skin lesion classification on the MCVSLD dataset. Its high accuracy, precision, and recall across all classes, particularly on critical ones, validate the strength of combining a domain-specific CNN with a state-of-the-art vision transformer. While the Custom CNN + MobileNetV2 were close seconds, the Custom CNN and fine-tuned VGG16 and Custom CNN + Swin Transformer models has proven to be a highly effective third option, demonstrating a great balance of performance and computational efficiency. These findings provide a robust foundation for further development and clinical validation of this diagnostic tool.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Introduction

The creation and use of an advanced AI model for medical diagnosis can be used to extract knowledge that goes beyond simple development and validation curves. This chapter, too, touches on the wider society-ethical-environmental aspects and things we as developers also need to have in mind when creating new technology.

5.2 Societal Impact and Access to Healthcare

The greatest public good we in dermatology could derive from this would be if we were able to democratize who gets access to expert-level diagnosis by a dermatologist. In regions where there is a lack of access to trained providers, an AI system that can work on mobile phones could serve as a potentially life-saving triage tool for frontline health workers and patients to guide informed decisions about whether or not to seek expert care. This could help identify life-threatening illnesses — including melanoma — earlier, potentially saving patients' lives and lightening the load for overstretched health care systems. The technology is also a potential equalizer in health care disparities, helping to get specialized expertise to isolated or underserved areas.

But the rapid advancement of this technology has its downsides. The "digital divide" can also widen existing health disparities if certain groups are not able to access enough digital infrastructure, cannot afford smartphones or lack the needed digital literacy for mutually beneficial use of such technology. Moreover, excessive dependence on AI with no human surveillance may result in misinterpretation of rare or atypical deformities which are poorly presented within the training data.

5.3 Ethical Considerations and Data Responsibility

There are ethical dimensions to such work, particularly around data privacy, algorithmic bias and accountability. Management of large sets of patient images and metadata are also needed by authors cited in the literature review. Both the collecting and usage of the data has to be recorded by patient consent, policy must be made for how to keep and use this information, and when anonymization is performed.

Algorithmic bias is an important ethical concern. “If training data is not comprehensive, then model might perform worse in skin types or conditions or population that it hasn’t seen enough. This could result in a system that’s less accurate for some racial or ethnic groups, which is what some would term technological bias. It’s up to us as developers to make sure that it is only patient-based splits, metadata is fully sanitised and really it provides a cautious class-aware augmentation against these biases so the model can be fair and robust for everyone. ADeep learning models are also “black box,” says he, posing the question of how to hold the technology accountable. In a misdiagnosis, we must also be able to comprehend as generically as possible what the AI was saying so that the technology simply enhances but does not replace clinical judgment of a doctor.

5.4 Environmental and Sustainability Implications

Training and running large-scale deep learning models has a non-negligible environmental cost. Carbon emissions are also produced by energy consumed by data centers and hardware for high performance computing. Models are getting bigger and the complexity is increasing which also means a bigger environmental impact. Research and practitioners should factor this cost into their work, and investigate more energy-efficient arch itectures, model compression techniques, and the use of smaller, task -specific models where possible.

On the other hand, there is a strong argument for the sustainability of this technology. Providing remote diagnostics and avoiding non-essential clinic visits could result in considerable savings in travel-related carbon emissions. In addition, a reduction in patient

transport can have a positive environmental impact due to the minimisation of long distance travel between a patient and his/her healthcare resources.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Summary of Findings

This thesis dealt with an urgent requirement of an accurate, cost effective and user-friendly diagnostic modality for viral skin lesions. This paper is based on extensive research, from an extensive literature review to the implementation and deployment of a working mobile application. The primary aim was to determine the best hybrid deep learning architecture that achieves both excellent classification capabilities and on-device performance feasibility.

After performing a thorough comparative analysis between five different hybrids, the study has produced a rather trivial, yet significant result: that the Custom CNN + EfficientNetB0 is the most suitable architecture. This representation obtained a high overall accuracy of 99%, confirming not only the assumption that it is possible to reinforce the classifier performance by using a feature-based fusion approach, but also outperforming performance of other models on the task. With computational efficiency (5.6 MB model, inference time 50ms) that proved stunning levels of performance do not necessarily require a compromise in practicality. The fact that this model can be translated and to a certain degree integrated as well, in a mobile application, was in itself the direct answer of the core research question about on-device usage, showing that such an end-to-end solution is actually feasible.

6.2 Research Contributions

This work makes several significant contributions to the fields of deep learning and AI-assisted medicine:

- **Systematic Comparative Analysis:** I conducted the first systematic, head-to-head comparison of a diverse range of hybrid deep learning models for the specific task of multi-class viral skin lesion classification on a unified dataset. This provides a clear, data-driven recommendation for future research in this area.

- **End-to-End Practical Solution:** Beyond theoretical model performance, this thesis provides a complete, end-to-end pipeline that transforms a trained model into a practical, real-world application. This includes the crucial steps of model optimization and on-device deployment, which are often overlooked in academic studies.
- **Privacy-Preserving and Accessible Framework:** By designing the application to perform all inference locally on the device, this research provides a scalable framework for creating AI tools that prioritize user privacy and can function reliably in low-resource settings without internet connectivity.

6.3 Future Work

The success of this research provides a strong foundation for future advancements. The following are suggested directions for continued work:

- **Enhancing Dataset Diversity:** The generalizability of the model can be significantly improved by training it on a more diverse dataset that includes images from a wider range of skin tones, lighting conditions, and geographic locations. This would mitigate potential algorithmic biases and enhance the model's performance in varied clinical settings.
- **Multi-modal Diagnostics:** A powerful next step is to integrate a multi-modal approach. Future work could explore incorporating clinical metadata, such as patient age, gender, and reported symptoms, into the model's decision-making process. This fusion of image features and patient data has the potential to provide a more nuanced and accurate diagnosis.
- **Advanced Explainability:** To build greater clinical trust, future research should focus on implementing and evaluating advanced explainability techniques, such as Grad-CAM++, to provide visual insights into the model's predictions. This would help clinicians understand *why* the model made a certain decision, transforming it from a "black box" to a collaborative diagnostic assistant.

- **Clinical Validation:** The ultimate validation of this work would be a formal clinical trial to test the mobile application's performance in a real-world setting. A formal trial would provide empirical evidence of the tool's clinical efficacy and safety.

6.4 Concluding Remarks

This thesis has proven that a hybrid deep learning model carefully designed and finely-tuned provide an effective and efficient means for the multi-class classification of viral skin lesions. By overcoming the hurdles of visual ambiguity and computational power, we have devised an answer which is not only a resounding research triumph but also a cost-effective, deployable solution that will change healthcare service delivery and diagnostic efficiency in dramatic ways. The insights from this work contribute to the vision of AI-assisted tools as commonplace in global public health.

REFERENCES

[1]. Gulzar, Y., Agarwal, S., Soomro, S., Kandpal, M., Turaev, S., & Choo, W. (2025). Next-generation approach to skin disorder prediction employing hybrid deep transfer learning. Frontiers in Big Data, 8, 1503883. https://doi.org/10.3389/fdata.2025.1503883

[2]. Agarwal, S., & Mahto, A. K. (2025). Skin cancer classification: Hybrid CNN-Transformer models with KAN-based fusion. arXiv preprint arXiv:2508.12484. https://arxiv.org/abs/2508.12484

[3]. Jabber, A. A., Shadeed, G. A., Salim, N., & Dibs, H. (2025). Automated skin lesion diagnosis and classification using K-means, LAB-color-space segmentation and deep learning. Operations Research Forum, 6, 43. https://doi.org/10.1007/s43069-025-00430-3

[4]. Reddy, D. S., Padmaja, N., Kumar, J. N., Gutam, B. G., Kumar, S. R., & Kumar, M. S. (2024). Enhanced skin cancer classification with an Attention-CNN-Transformer model. Frontiers in Health Informatics, 13(3), 809. https://doi.org/10.52547/fhi.13.3.809

[5]. AIP Advances. (2025). Efficient classification of skin disease using deep learning technique (EfficientNet-B3). AIP Advances, 3257, 020135. https://doi.org/10.1063/5.0264862

[6]. Zhang, X., et al. (2025). A novel hybrid ConvNeXt-based approach for enhanced skin lesion classification. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2025.125512

[7]. Khan, M. A., Rastogi, D., Johri, P., Al-Taani, A., Baghela, V. S., & Kumud. (2025). Hybrid deep CNN model for multi-class classification of skin lesion. Neural Computing and Applications, 37, 19479–19499. https://doi.org/10.1007/s00521-025-11409-w

[8]. Wang, X., Zhang, X., & Wang, X. (2025). Deep skin lesion segmentation with Transformer-CNN fusion: Toward intelligent skin cancer analysis. arXiv preprint arXiv:2508.14509. https://arxiv.org/abs/2508.14509

[9]. Hasan, M. Z., & Rifat, F. Y. (2025). Hybrid ensemble of segmentation-assisted classification and GBDT for skin cancer detection. arXiv preprint arXiv:2506.03420. https://arxiv.org/abs/2506.03420

[10]. Manivannan, S. (2025). Semi-supervised learning with online knowledge distillation for skin lesion classification. arXiv preprint arXiv:2508.11511. https://arxiv.org/abs/2508.11511

[11]. Akter, M., Khatun, R., Talukder, M. A., Islam, M. M., & Uddin, M. A. (2024). An integrated deep learning model for skin cancer detection using hybrid feature fusion technique. arXiv preprint arXiv:2410.14489. https://arxiv.org/abs/2410.14489

[12]. Pradeepa, R., Punitha, V., & Selvi, R. S. (2024). Deep learning algorithms for skin disease classification. Journal of Innovative Image Processing, 6(2), 84–95. https://irojournals.com/iroiip/article/view/6/2/1

[13]. Tejasri, N. L. S. V., et al. (2025). Deep learning strategies for skin disease classification using VGG16 and EfficientNet. In Lecture Notes in Computer Science (LNCS) (pp. 225–236). Springer. https://doi.org/10.1007/978-3-031-77075-3_18

[14]. Kumar, V., et al. (2025). Skin cancer classification using CNN with transformer layer integration. In Lecture Notes in Computer Science (LNCS) (pp. 115–126). Springer. https://doi.org/10.1007/978-3-031-77075-3_13

[15]. Agarwal, S., et al. (2025). CNN classification for four types of skin cancer using HAM10000 dataset. International Journal for Multidisciplinary Research (IJFMR), 7(3), 43814. https://www.ijfmr.com/papers/2025/3/43814.pdf

[16]. Gupta, P., Nirmal, J., & Mehendale, N. (2024). Custom CNN architectures for skin disease classification: Binary and multi-class performance. Multimedia Tools and Applications, 84, 32505–32532. https://doi.org/10.1007/s11042-024-20503-5

[17]. Pandey, A., Teja, M. S., Sahare, P., Kamble, V., Parate, M., & Hashmi, M. F. (2024). Skin cancer classification using non-local means denoising and sparse dictionary learning-based CNN. Journal of Electrical Systems and Information Technology, 11, 36. https://doi.org/10.1186/s43067-024-00162-0

[18]. Shukla, M. M., Tripathi, B. K., Dwivedi, T., Tripathi, A., & Chaurasia, B. K. (2024). A hybrid CNN with transfer learning for skin cancer disease detection. Medical & Biological Engineering & Computing, 62, 3057–3071. https://doi.org/10.1007/s11517-024-03115-x

[19]. Mohan, J., Sivasubramanian, A., Sowmya, V., & Vinayakumar, R. (2024). Enhancing skin disease classification leveraging Transformer-based deep learning architectures and explainable AI. arXiv preprint arXiv:2407.14757. https://arxiv.org/abs/2407.14757

[20]. Himel, G. M. S., Islam, M. M., Al-Aff, K. H., Karim, S. I., & Sikder, M. K. U. (2024). Skin cancer segmentation and classification using Vision Transformer for automatic analysis in dermatoscopy-based non-invasive digital system. arXiv preprint arXiv:2401.04746. https://arxiv.org/abs/2401.04746

[21]. Ahmed, Kazi Rifat; Rashid, Md. Mazbaur ; Sarkar, Suprove Chandra ; Islam, Mujahidul (2025), “Multi-Class Viral Skin Lesion Dataset (MCVSLD)”, Mendeley Data, V1, doi: 10.17632/dfztdtfsxz.1

242-25-042

ORIGINALITY REPORT

12%	7%	8%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	1%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	1%
4	www.mdpi.com Internet Source	1%
5	Submitted to Florida National University Student Paper	<1%
6	Burhanettin Ozdemir, Fethi Sermet, Ishak Pacal. "Attention-enhanced ConvNeXt for accurate, efficient, and interpretable crack detection", Expert Systems with Applications, 2025 Publication	<1%
7	Submitted to University of Westminster Student Paper	<1%
8	"Advances in Smart Computing and Applications", Springer Science and Business Media LLC, 2026 Publication	<1%