Efficient and Explainable Skin Lesion Classification: A Single ConvNeXt-Large Surpassing ConvNeXt Ensembles with Grad-CAM

By

Md. Omar Farooq
20201154
Md Azmain Khan Alve
20301399
Kaniz Fatema Kanta
22201671
Md. Rayhan Sarker
20201165
Md Shahadat Hossain
20201181

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
June 2025

© 2025. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Omar Farooq
20201154

Md Azmain Khan Alve
20301399

Kaniz Fatema Kanta
22201671

Md. Rayhan Sarker
20201165

Md Shahadat Hossain
20201181

Approval

The thesis titled "Efficient and Explainable Skin Lesion Classification: A Single ConvNeXt-Large Surpassing ConvNeXt Ensembles with Grad-CAM" submitted by

1. Md. Omar Farooq (20201154)
2. Md. Azmain Khan Alve (20301399)
3. Kaniz Fatema Kanta (22201671)
4. Md. Rayhan Sarker (20201165)
5. Md. Shahadat Hossain (20201181)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on June 20, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Automatic classification of skin lesions from dermoscopic images plays a crucial role in early cancer detection. However, this task remains challenging due to visual similarities between lesion types, class imbalance, and image artifacts. While recent deep learning approaches, including ConvNeXt-based ensembles, have demonstrated impressive classification accuracy, their clinical reliability is hindered by uneven performance across classes, especially lower sensitivity in detecting high-risk conditions. In this study, we propose an efficient and interpretable framework based on a single ConvNeXt-Large model,trained and evaluated on the ISIC 2019 skin lesion dataset. Through systematic empirical optimization, we achieved a balanced performance that surpasses ensemble methods in clinically critical metrics, attaining a sensitivity of 86.27% and a specificity of 98.08% on the validation set. This was accomplished without compromising overall accuracy, while significantly reducing computational complexity compared to ensemble architectures. To enhance transparency and clinical trust, we further integrated Grad-CAM visual explanations, which reveal that the model consistently focuses on medically relevant lesion regions. Overall, this work contributes a practical, lightweight, and trustworthy deep learning solution for skin lesion classification that balances accuracy, efficiency, and interpretability—making it more suitable for real-world deployment.

Keywords: Deep Learning, Skin Disease Classification,ConvNext Large, ConvNext ensemble, Explainable AI, Grad-CAM, ISIC 2019 Dataset, Medical Image Analysis.

Acknowledgement

We want to thank the following people and organizations from the bottom of our hearts for all the help and support they gave us while we were working on this thesis. We would like to thank **Dr. Ashraful Alam**, our thesis **supervisor**, from the bottom of our hearts for always being there for us, giving us helpful feedback, and being patient with us during our research journey. Your knowledge and help have been very helpful in deciding how this study will go. The authors would like to thank the instructors and staff at **BRAC University** for giving them the tools, space, and support they needed to do this study. I would not have been able to finish this thesis without the help and support of these people and organizations. Thank you for being a part of this academic project.

1. **Md. Omar Farooq**

2. **Md. Azmain Khan Alve**

3. **Kaniz Fatema Kanta**

4. **Md. Rayhan Sarker**

5. **Md. Shahadat Hossain**

Table of Contents

Declaration i

Approval ii

Abstract iii

Acknowledgment iv

Table of Contents v

List of Figures vii

List of Tables viii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Problem Statement . . . 3
- 1.3 Research Objectives . . . 5

2 Literature Review 6
- 2.1 Optimized Deep Learning Architectures for Skin Lesion Classification 6
- 2.2 Explainable AI for Transparent and Trustworthy Diagnostic Decisions 8
- 2.3 ISIC Dataset Challenges and Solutions for Multiclass Skin Lesion Modeling . . . 10
- 2.4 Clinical Integration and Real-World Applications . . . 12

3 Methodology 14
- 3.1 Dataset Preprocessing And Augmentation . . . 14
  - 3.1.1 Dataset Description . . . 14
  - 3.1.2 Data Visualization . . . 14
  - 3.1.3 Class Distribution Analysis . . . 15
  - 3.1.4 Test Set Integration for Dataset Expansion . . . 16
  - 3.1.5 Dataset Splitting Strategy . . . 17
  - 3.1.6 Data Augmentation for Class Balance . . . 17
- 3.2 Model Architecture and Image Flow . . . 19
- 3.3 The Iterative Training Pipeline: A Step-by-Step Optimization Journey 21
  - 3.3.1 Initial Setup and Early Observations (Epoch 1-7) . . . 21
  - 3.3.2 CoarseDropout Implementation, Augmentation Enhancements, and Focal Loss Transition (Epoch 8–29) . . . 22

3.3.3 Comprehensive Augmentation, Dynamic Class Weighting, and Cosine Annealing Strategy (Epoch 30–56) 23
3.3.4 Further Class Weight Tuning and Focal Loss Gamma Refinement (Epoch 56-76) 24
3.3.5 Class Weight Recalibration and Sensitivity Stabilization (Epoch 77–106) 24
3.3.6 Final Class Weighting Strategy and Peak Model Performance (Epoch 105–130) 25
3.4 Explainable AI Implementation (Grad-CAM) 26

4 Result and Analysis **28**
4.1 Evaluation Metrics 28
4.2 Evaluation of Validation and Test Results 29
4.2.1 Final Result Summary 29
4.2.2 Classification Results 29
4.2.3 Loss Analysis 30
4.3 Class-wise Analysis and Comparative Discussion 31
4.3.1 Comparative Analysis of the Final Results 33
4.4 Explainability with Grad-CAM Visualization 34
4.4.1 Heat Map Visualization 34
4.4.2 Interpretation and Clinical Impact 35

5 Conclusion **37**

Bibliography **40**

List of Figures

3.1 Data visualization from each class 15
3.2 A pie chart illustrating the initial class distribution of the 25,331 images 16
3.3 Original vs Augmented Images 18
3.4 Before and after count of data Augmentation 19
3.5 Model Architecture & Image Flow Diagram 21

4.1 Train Loss vs. Validation Loss over 130 epochs 30
4.2 Grad-CAM heatmaps highlighting lesion-focused attention in correct and incorrect predictions. 35

List of Tables

3.1 Test Set Integration for Dataset Expansion 16
3.2 Train, Validation, and Test Set Distribution by Class 17

4.1 Summary and Comparison of Validation and Test Results 29
4.2 Classification Report of the Test Set 30
4.3 ConvNeXt Large models vs existing models (Performance comparison) 33
4.4 Comparison of ConvNext Large Model vs. ConvNeXt Ensemble . . . 34

Chapter 1

Introduction

1.1 Background

Early and accurate diagnosis is crucial for managing skin lesions, from common benign conditions to life-threatening melanomas, in order to improve patient outcomes around the world [4]. Skin lesions are areas of skin that don't look like the rest of the skin around them. They can be caused by injuries, sun exposure, infections, or problems with the immune system. Most of the skin lesions, for example, birthmarks, moles, skin tags, and acne, are not cancerous and are not harmful (benign). However, they can sometimes be a sign of a more serious condition, such as an infection, an autoimmune disease, or, most importantly, skin cancer [30].

Skin cancer is one of the most common types of cancer in the world, and the number of cases and deaths from it rises every year.According to the GLOBOCAN 2022 data, melanoma was the 17th most common cancer, with an estimated 331,722 people diagnosed and about 58,667 deaths worldwide[28]. This growing number puts a lot of stress on healthcare systems around the world. This growing number puts a lot of stress on healthcare systems around the world. Finding skin cancer, especially melanoma, early is a good goal because it can save lives, lower the number of sick people, and maybe even lower healthcare costs. For example, finding melanoma early makes a big difference in the outcome; for stage IA, the five-year survival rate is 98%. It's very important to stress early diagnosis because advanced melanoma can spread to other organs, making treatment more complicated and expensive and lowering survival rates. So, one of the best path to fight the growing problem of skin cancer is to improve diagnostic tools as soon as possible.

It is naturally hard to visually assess these lesions because they often look very different from one another and are hard to interpret clinically [29]. Dermoscopy is a non-invasive imaging method that has greatly improved dermatologists' ability to see subsurface skin features. This gives them important information that makes diagnoses more accurate than just looking at the skin. Digital dermoscopy is a more advanced type that uses computerized imaging to give high-resolution, enlarged pictures. This lets doctors closely look at lesions and keep an eye on them over time for small changes [29]. This method helps find things that can't be seen with eyesight alone, like blood vessels and pigment networks. This makes it more straightforward to diagnose skin cancers.

Dermoscopy technology has come a long way over the years . Digital dermoscopy is a more advanced type that uses computers to make high-resolution, enlarged images. This lets doctors closely look at lesions and keep an eye on them over time to see if there are any small changes [29]. Another improvement is the use of polarized light, which cuts out surface reflections to make it easier to see deeper structures and find malignant features . Reflectance Confocal Microscopy (RCM) is another way to see skin cells without cutting them open. It's like a virtual biopsy. These advances in technology, along with the use of AI algorithms, make dermatopathology more accurate and based on data [29].

Medical image analysis is complicated by nature, especially when it comes to visual diagnosis. This is true in many medical fields, not just dermatology. These problems come from things like small differences in pathology, the need for highly specialized expert interpretation, and often, a lack of data or balance [4]. In this context, deep learning has completely changed the way medical images are analyzed. It now allows for automated feature extraction and advanced pattern recognition in various fields, such as radiology, pathology, and ophthalmology. Among the many types of deep learning architectures, Convolutional Neural Networks (CNNs) have always been the most important for tasks that involve images. This is because it can learn hierarchical visual representations directly from raw pixel data [4]. The fact that CNNs are widely used in medical diagnostics shows how well they can handle complicated visual data for tasks like classification and segmentation [4]. More recently, digital dermoscopy systems have added artificial intelligence (AI) and machine learning algorithms. These have greatly improved the accuracy of diagnoses by making it easier to find cancers and speeding up and making the analysis of dermoscopic images more reliable [29]. AI algorithms, especially those that use deep learning, are very good at recognizing skin images and telling the difference between different types of skin lesions. Their diagnostic accuracy is similar to that of dermatologists [4]. This integration helps find patterns linked to different skin lesions, likely melanoma and non-melanoma skin cancers. This makes diagnoses more accurate and decreases human error [29].

This thesis presents a deep learning-based architecture developed in an optimized single ConvNeXt-Large model. It also includes Grad-CAM Explainability to create a skin disease classification system that can be understood. The framework is meant to provide better and more balanced diagnostic performance than complex ensemble-based methods, while still being highly transparent and straightforward to understand. This makes sure that AI-based diagnostic systems in dermatology are both clinically useful and reliable.

1.2 Problem Statement

Millions of people around the world have skin cancer, from common skin problems to deadly melanomas. Immediate medical care and early diagnosis of deadly cancers like melanoma are key to successful treatment and greatly affect the patient's chances of survival. Modern diagnostic methods mostly depend on dermatologists' visual assessments [29]. Even though early detection is very important in clinical practice, current traditional methods rely a lot on the skills of individual dermatologists, which can lead to inconsistent results and mistakes because they are subjective and take a long time [13].

Human error is a major risk in dermatology since diagnosis is mostly dependent on experience-based assessment, which can result in problems including attention bias, fatigue, and inconsistent expert abilities[13]. When comparing visually identical disorders, such as melanoma vs seborrheic keratosis, dermatologists' diagnostic skills can differ greatly between highly experienced doctors and less experienced specialists [13]. Inconsistent diagnosis rates may arise from this difference in experience, raising the possibility of postponing the identification of skin cancer and permitting it to evolve into significant treatment challenges.

A global lack of qualified dermatologists significantly impedes patients' ability to obtain timely medical assessments, particularly in rural and limited-income areas, leading to extensive wait times for specialist examinations[4]. The escalating incidence of dermatological diseases further burdens already overloaded healthcare facilities, increasing demand for medical services. In cases requiring urgent medical intervention, the absence of readily available specialists heightens the risk of disease complications for patients [4].

Determining skin disease diagnoses is challenging due to the various patterns of shape, size, and color variations displayed by skin lesions [29]. Dermatologists, regardless of experience level, often encounter difficulties differentiating between cancerous and non-cancerous lesions, frequently leading to the use of invasive diagnostic procedures, such as biopsies. However, the clinical utility of biopsies is limited by their requirement for specialized facilities, their potential to cause patient discomfort, and their infeasibility in resource-constrained healthcare environments [29]. Consequently, there is an urgent requirement for automated diagnostic systems that can perform non-invasive testing, thereby eliminating the need for subjective analysis and invasive medical procedures.

Convolutional Neural Networks (CNNs) have been important for tasks that involve images, but traditional CNN architectures often have trouble capturing long-range dependencies and understanding the global context because their receptive fields and parameters are fixed. Vision Transformers (ViTs) fix these problems with self-attention mechanisms, but they usually make things harder to compute [19]. It is hard to find an architecture that can combine the efficiency of CNNs with the ability of transformers to understand the whole picture. The ConvNeXt architecture, for example, tries to do this by adding transformer-like configurations to a CNN framework while keeping efficiency [19].

Even though deep learning has come a long way, models still don't always work well across all classes, especially when trained on datasets that aren't balanced. To fix this, researchers have looked into ConvNeXt-based ensemble models, which can improve accuracy but come at a high cost [13]. These ensemble methods make things much more complicated: they take a long time to train, need a lot of computational power, and are hard to use in real-world clinical settings. When you coordinate multiple models, it takes longer to make inferences and makes them harder to understand, which makes them less useful for everyday use [13]. This makes it clear that we need a single, optimized model that can perform well and evenly across all eight classes while still being efficient, easy to understand, and useful in a clinical setting.

Combining deep learning and artificial intelligence (AI) presents a viable way to classify skin diseases in order to overcome these complex challenges. Convolutional neural networks (CNNs) are widely used in medical imaging, but their clinical deployment often fails due to their opaque nature, which is known as the "black box" problem [13]. Until they have a clear understanding of how these systems make their clinical judgments, healthcare professionals usually clash with the use of AI diagnostic models, which interferes with trust and prevents wider adoption. The broad and reliable use of AI-driven diagnostic tools in dermatology is severely hampered by this lack of transparency.

1.3 Research Objectives

The purpose of this thesis is to develop an effective and interpretable deep learning framework that can achieve robust, class-balanced diagnostic performance in order to address the ongoing limitations in skin lesion classification. The study is especially concerned with addressing the disadvantages of ensemble-based ConvNeXt models, which frequently lack sensitivity and class-level fairness in spite of their high overall accuracy. The ISIC 2019 dataset, a sizable, real-world, but extremely unbalanced dermoscopic image dataset with eight different lesion class—including benign and malignant cases—serves as the foundation for our work. Its difficult nature, characterized by rare critical classes (like SCC, AK) and dominant majority classes (like Nevus), calls for precision-focused solutions that go beyond exaggerated accuracy metrics. The specific objectives include:

- Developing an AI-based diagnostic system that can help clinicians and reach underserved populations where dermatological expertise is limited will ultimately prevent delayed diagnosis and lower health risks while also improving accessible and early skin cancer detection.

- To use the ConvNeXt-Large architecture to create a single-model diagnostic system that maintains high accuracy and clinical reliability while doing away with the need for resource-intensive ensemble models.

- By using adaptive loss weighting and advanced data augmentation, it is possible to address the crucial amount of class imbalance in the ISIC 2019 dataset and improve classification performance for both common and uncommon lesion types.

- To maximize sensitivity and specificity while maintaining generalizability by optimizing the model using a feedback-driven empirical approach that makes use of mixed precision training, learning rate scheduling, staged fine-tuning, and albumentations-based augmentations.

- To improve clinical trust and transparency by using Explainable AI (XAI), namely Grad-CAM, which aligns the model's focus with dermatological reasoning and allows for visual interpretation of the model's decisions.

- To assess the clinical readiness of the model by means of rigorous testing on unseen and validation test sets, utilizing key diagnostic metrics (accuracy, sensitivity, specificity, precision, and F1-score), and contrasting it with the most advanced benchmarks, such as ensembles based on ConvNeXt.

- To provide a useful, interpretable, and computationally effective diagnostic solution that can be implemented without sacrificing diagnostic safety at different levels of healthcare infrastructure, from remote telemedicine setups to urban hospitals.

Chapter 2

Literature Review

Deep learning and explainable AI (XAI) have made a big difference in the field of skin lesion classification by making models that are both accurate for diagnosis and easy to understand in a clinical setting. This literature review examines studies on how deep learning architectures have changed over time, especially how they have moved toward transfer learning and ensemble techniques for classifying dermoscopic images. It also looks at how XAI methods like Grad-CAM help make models more open and trustworthy in clinical settings. There are still problems with benchmark datasets like ISIC, such as severe class imbalance and small similarities between classes that make it hard to generalize. Finally, the review talks about attempts to bring AI-powered diagnostic systems into real-world medical facilities where resources are limited. These focus areas form the basis of this study and show the important gaps that our proposed ConvNeXt-Large framework will fill.

2.1 Optimized Deep Learning Architectures for Skin Lesion Classification

The study by Kablan and Ayas [27] introduces a novel ensemble-based approach for multi-class skin lesion classification, leveraging the distinct capabilities of various ConvNeXt models. This research underscores the critical importance of automated dermoscopy image analysis in facilitating the early and precise diagnosis of skin cancer, a key factor in improving patient prognosis. The authors capitalize on the advanced capabilities of deep learning-based methods, which have recently demonstrated superior performance in skin lesion classification tasks compared to conventional machine learning techniques. Their methodology meticulously investigates different versions of pre-trained and fine-tuned ConvNeXt models—namely Tiny, Small, Base, and Large—for robust classification. Crucially, they address persistent challenges such as class imbalance through the implementation of a weighted loss function and enhance model generalization via comprehensive data augmentation techniques. The application of an ensemble method, combining the outputs of these particular models, yielded a demonstrably strong diagnostic performance, characterized by a sensitivity of 84.2% and a specificity of 97.9%, thereby highlighting the efficacy of integrating diverse model strengths for improved accuracy in complex medical imaging tasks

Dhankar et al.'s [22] study uses Convolutional Neural Networks (CNNs) and Random Forest (RF) classifiers to find skin diseases early. The goal is to provide automated healthcare solutions that don't require surgery. It did this by using a CNN model that got 88.5% accuracy overall, but got better in some cases, reaching 90.02%. It used Histograms of Oriented Gradients (HOG) to extract features and data to fix class imbalance. We got some good results, but we still have some problems because the dataset is small and there are too few examples of some classes, which makes it hard to fit the data. Also, different devices in different regions don't always work together.

The study by Akyeramfo-Sam et al.[6] similarly highlights a web-based diagnosis system for skin diseases by Convolutional Neural Networks (CNNs), with the emphasis on three most common conditions: atopic dermatitis, acne vulgaris and scabies. This system could classify atopic dermatitis, acne vulgaris and scabies images with 88%, 85% and 84.7% accuracies using the TensorFlow framework while providing a prediction in as short as time of just over one milliseconds. This solution is a response to the shortage of dermatologic resources in Ghana and innovates with use of an AI-powered method that outperforms human-based methods in both speed and accuracy for diagnosis. That said, the study was not without its limitations — there were a few challenges to address in order to better generalize results across different types and severities of skin. In the future work we propose to integrate hybrid algorithms and batch image processing for an increased performance, as well validate scaling up of this method towards wider healthcare applications. Results suggest it is the perfect way to improve aid in medical diagnostics, especially for countries with constricted healthcare infrastructure.

Allugunti's paper[16] suggests using deep learning to sort skin diseases, with a focus on melanomas as the most dangerous type of skin cancer that can be found early on. Using CNNs, this study sorted different types of melanoma based on demographic data from DermNet NZ, with an accuracy rate of 88.83%. The study's results show that CNN is better at handling complex image data than traditional classifiers like Decision Trees and Random Forests. However, the dataset's limitations included restrictions and uneven representation across different types of melanoma. The study shows that using automated tools to find melanoma early is important for improving patient outcomes. This helps dermatologists by giving them a more accurate initial profile of the suspected case.

Chowdhury et al.[23] did a study that used four different Convolutional Neural Network (CNN) architectures—EfficientNetV2B3, EfficientNetV2s, Inception Net V3, and Densenet121—on the HAM10000 dataset. Each architecture got an accuracy of 83%, 86%, 84%, and 88%, respectively. The research shows that an ensemble method can get the average ensemble up to 90% right, and a K-Nearest Neighbors (KNN)-based ensemble got it up to 92%. Explainable AI (XAI) methods were also used to make the model easier for doctors to understand and to build trust and transparency. In conclusion, this study shows how important it is to combine deep learning and XAI to make strong, accurate tools for diagnosing skin cancer.

A deep learning methodology designed by Kassem et al.[9] brought multi-class skin

lesion identification to the ISIC 2019 data repository with eight distinct types of skin lesions. This research incorporated transfer learning methods through which pre-trained onal Neural Network (CNN) models including ResNet-50, InceptionV3 as well as MobileNetV2 contributed to accuracy improvements. The researchers handled class imbalance by implementing data augmentation operations that included flipping and rotation along with contrast modifications. Their most successful model achieved dermatological image classification with an exceptional accuracy level of 94.67%. The research study established that feature extraction in combination with hyperparameter tuning works as key components for enhancing model robustness levels. The research indicates deep learning technology provides dependable detection systems for skin diseases thereby improving medical diagnosis together with expert dermatological decisions.

One of the most detailed overviews of the deep learning applications in medical image analysis has been produced by Litjens et al. [1] and includes more than 300 papers on a variety of imaging modalities, including MRI, CT, ultrasound, and dermoscopy. The article concentrates on the fact that Convolutional Neural Networks (CNNs) have substituted using hand-crafted techniques by using a data-driven approach to learn features directly and are performing better in tasks such as The detection of tumors in organs, segmentation of organs, and classification of lesions. AlexNet, VGG, and U-Net, are popular models that were considered successful in the early years. Nevertheless, the authors uncover pinpointed challenges as well such as the lack of labeled data and data imbalance and lack of model interpretability that can serve as hindrances to the adoption of machine learning models in clinical settings. To overcome this, the paper suggests the use of data augmentation; transfer learning, and explainability methods such as saliency maps and Grad-CAM methods. It also emphasizes a necessity of more annotated data, benchmarks to allow fair comparisons, and readiness to real-world conditions. The fact is that this work became a thread on which future researchers could work, providing specific areas where deep learning can be applied in the health industry, as well as the limitations of this area

2.2 Explainable AI for Transparent and Trustworthy Diagnostic Decisions

Singh, Sengupta, and Lakshminarayanan [11] survey the importance of explainable deep learning (XAI) models in medical image analysis by increasing trustworthiness for clinicians as well as patients. The research emphasizes that explainability is crucial for guaranteeing transparency, legal compliance (such as GDPR), and successful clinical deployment, even though deep learning has shown outstanding performance in fields like Alzheimer's detection and retinal disease diagnosis. To illustrate model decisions, it divides XAI techniques into model-specific and model-agnostic approaches, emphasizing attribution-based techniques like GradCAM and SHAP.The study's conclusion highlights the necessity of additional explainability developments to support trustworthy AI integration in healthcare and guarantee that physicians can comprehend and have faith in the diagnostic insights these models offer.

Gohel, Singh, and Mohanty[14] talk about Explainable Artificial Intelligence (XAI) in their paper. They say that XAI is important for healthcare, defense, and autonomous systems. Black box AI models don't explain why they make decisions, so no one knows why they do. XAI wants to fix this by using tools like LIME, SHAP, and Layerwise Relevance Propagation (LRP) that let users look into how a certain answer was reached. These are the things that make trust possible in high-stakes situations like medical diagnostics or the legal system. The authors stress the need to find a balance between performance and interpretability for practical AI deployment. They also suggest that future research should focus on making XAI techniques more efficient so they can be used more widely in the real world.

This study by Nigar et al.[20] uses Explainable Artificial Intelligence (XAI) methods to make deep learning models for classifying skin lesions easier to understand and more reliable. The authors used the ISIC 2019 dataset to pre-train the ResNet-18 model, which had an accuracy of 94.4%, a precision of 93.57% , a recall of 94.01%, and an F1 score of 94.45%. The authors added LIME (Local Interpretable Model-Agnostic Explanations) to make it easier to understand. To make it more useful in a clinical setting, they also added a visual explanation for the model's prediction. Using coupled deep learning with explainable AI not only makes diagnoses more accurate, but it also helps build trust by being used in decision-making processes. This will lead to widespread use in the healthcare field.

Ahmad et al.'s[21] study suggested a way to recognize multiple types of skin lesions by combining deep learning with Explainable AI (XAI). The method used the Xception and ShuffleNet models to get features. On the other hand, the Butterfly Optimization Algorithm (BOA) is used to choose the best features to improve the model's performance. The system did very well, getting 99.3% correct on the HAM10000 dataset and 91.5% correct on the ISIC2018 dataset. The authors used GradCAM-based visualization to highlight important parts of input images in order to make them easier to understand, especially for use in clinic centers. This study can help make deep learning-based models easier to understand and more reliable for automatically classifying skin lesions to help doctors make decisions.

Radiah et al.'s[24] study shows how to use Vision Transformers (ViT) and Convolutional Neural Networks (CNN) together with Explainable AI Techniques to make Skin Cancer Detection easier to understand. Their tests showed that VGG-16 had the highest accuracy, at 82.14%, while ViT models had accuracies between 74.55% and 76.15%. GradCAM visualizations were used in the study to show which parts of the model were most important for making decisions. This made the decision-making process clear and easy to understand.The study included Grad- CAM visualizations to identify salient regions responsible for model decisions, thus making the decision-making process transparent and interpretable. It was all for transparency and wanting the system to be as reliable a tool in clinics so that der- matologists can check if its predictions were ever wrong, thus avoiding false positives—results turned out incorrect from machine learning. This transparency is created to make the system safer for clinical use, as it allows dermatological professionals to find out what alerts a model has messaged and manage its notifications appropriately so

patient confidence can be maintained by lowering false diagnosis counts.

The study by Badhon et al. [26] explore the use of Explainable AI (XAI) on Transfer Learning for multiple skin diseases detection such as acne, nail psoriasis and vitiligo. Along with VGG-16, VGG 19, as well as Xception a total of five CNN-based models were implemented. And the highest accuracy is 97.07% by VGG-16 on data augmentation training phase This points to the importance of balancing data through appropriate techniques in-order to perform better. They used Grad-CAM (gradient-weighted class activation mapping) visualizations to interpret the decisions of the model, pointing out relevant regions in images for interpretation and generalizability to iterative clinical use. According to the study, deep learning combined with XAI increases diagnostic accuracy, boosts confidence in AI-based technologies, and helps overcome data shortages.

2.3 ISIC Dataset Challenges and Solutions for Multiclass Skin Lesion Modeling

Through ISIC 2017 Challenge Codella et al[2]. established a framework to enhance melanoma detection by analyzing images of the skin. The challenge contained three main segments: lesion segmentation, dermoscopic feature detection and disease classification. Different participants used deep learning alongside ensemble methods when approaching these tasks. The challenge of standardization of data alongside tests made it possible to conduct fair evaluations between competing methods. The research findings showed that data augmentation together with transfer learning and multiple algorithms use improved system effectiveness. The benchmark provided tremendous advancement to automated skin lesion analysis through its creation of evaluation standards that united the research community efforts.

The combination of different Convolutional Neural Networks (CNNs) constructed an ensemble system for enhanced lesion segmentation in melanoma detection according to Al-Masni et al. [4]. Various CNN architectures within their method worked together to extract different types of features which improved segmentation precision. Multiple ensemble models achieved superior results than single CNNs because ensemble learning delivers effective performance in medical image evaluation. Research findings demonstrate that deep learning ensemble systems have strong potential for raising the precision along with dependability of automatic skin lesion evaluation algorithms.

Tschandl et al. [5] established the HAM10000 dataset by gathering 10,015 dermatoscopic images which encompass seven normal pigmented skin lesions. The HAM10000 dataset obtained its information from two sources that included the Department of Dermatology at the Medical University of Vienna and the skin cancer practice of Cliff Rosendahl in Australia. Specialists from the dermatology field performed detailed annotations on each image to achieve precise lesion categorization through high-quality labeling processes. HM10000 has evolved into a standard dataset for medical research because it facilitates essential machine learning algo-

rithm development for detecting skin lesions. The open availability of the database has resulted in significant progress for automated medical diagnosis of dermatological conditions by creating a large collection for AI model development and assessment.

Alzubaidi et al. [13] provide a detailed and more technical overview of deep learning and pay specific attention to Convolutional Neural Networks (CNNs) and their developments. The paper describes the approach of developing important architecture including ResNet, DenseNet, and EfficientNet in feature learning that address most typical challenges like vanishing gradients and ineffective parameters utilization. These innovations are particularly applicable in complicated image analysis procedures in the medical sphere. In the review, it also delves into larger notions such as data augmentation, transfer learning and optimization schemes to increase standardization of these models on small-sized medical training sets. Notably, the authors also discuss the issues that continue to restrict clinical application, i. e. the problems of class imbalance, overfitting, and opacity. They suggest hybrid models (e.g., CNN-RNN, CNN-GAN), and the inclusion of the interpretability measures, such as LIME and SHAP, to describe model choices. Future directions like the lightweight models of mobile health application and incorporation of domain knowledge to enhance reliability is also described in the paper. This first-of-its-kind piece of work is a technical roadmap of the researchers interested in the deep learning and medical imaging intersection.

Zhao et al. [12] suggest an enhanced way to solve the long-term problem affecting dermoscopic image classification of class imbalance and the scarcity of annotated data through the use of generative adversarial networks. In particular, through a new loss term to prioritize intra-class consistency and inter-class separability, the authors create an extended StyleGAN model specific in order to generate high-fidelity synthetic skin lesion images. These artificial images add to genuine information and feed into DenseNet-201 to group lesions into many classes. The research thoroughly tests this framework on ISIC 2019 benchmark, and obtains a significant increase in balanced multiclass accuracy. In order to solve the issue of data scarcity and improve diagnostic performance, the work contributes to the applied context of the combined usage of GAN-based data augmentation and cutting-edge CNN architectures.

The special concern of the paper by Iqbal I et al. [15] is the lightweight but effective deep convolutional neural network that is specifically aimed at the multi-class category of dermoscopic skin lesions. It is essential with regard to the deployment of their model architecture in real-world clinical settings where the presence of computational resources is relatively scarce. The focus on computational efficiency and decreased parameter complexity is an absolute necessity. It has been shown in the paper that competitive classification can achieve equivalent or worse compared to the performance of larger, resource-demanding networks on publicly released datasets and, therefore, offers an alternative to the former, without sacrificing accuracy. This work leads to the process of the balance of model complexity and performance the results available in the case of automated inspection of skin lesions systems

Ayas [17] proposes the use of Swin Transformer as a network architecture applied in

the field of skin lesion classification, which is another step of transitioning to using the transformer architecture in the field of medical images and such networks. The article outlines the effectiveness of the Swin Transformer to represent both local and global contextual features through the hierarchical attention adopted by the model since it plays a critical role in differentiating minor differences that characterize various types of lesions. By the method of addressing imbalance of classes using weighted cross-entropy loss, the model has been trained and validated on ISIC 2019 data to provide the state-of-the-art results as far as its accuracy, sensitivity, and specificity are concerned. The paper shows that the transformer can have a promising future in enhancing skin lesion classification and that a hybrid architecture will allow achieving future increases.

2.4 Clinical Integration and Real-World Applications

Agarwal et al.'s[25] study used the Gemini 1.0 Pro API to make the Skin Disease Prediction System work with images. This system was suggested along with the MobileNet CNN model. The system can be used on both onboard and offboard web-based platforms because it uses a lightweight and efficient MobileNet CNN, which is especially good for mobile apps. Users can upload pictures of their skin lesions with this app, which then analyzes them automatically and makes quick and accurate predictions about a number of conditions. This includes the Gemini 1.0 Pro API, which gives a more detailed description of each diagnosed condition, including its causes, symptoms, and treatments. This makes the app more interesting and easier to understand for users. Using machine learning models can make dermatological diagnostics cheaper when they are combined with easy-to-use interfaces, as shown in this study and others.

Li and Shen [3] proposed two deep learning frameworks to enhance automated skin lesion analysis: the Lesion Indexing Network (LIN) and the Lesion Feature Network (LFN). LIN integrates two fully convolutional residual networks to perform simultaneous lesion segmentation and classification, refining outputs with a Lesion Index Calculation Unit. LFN employs a straightforward convolutional neural network for dermoscopic feature extraction. Evaluated on the ISIC 2017 dataset, their models achieved Jaccard Index of 0.753 for segmentation, accuracy of 0.848 for feature extraction, and an area under the ROC curve of 0.912 for classification, demonstrating significant potential in improving melanoma detection.

Chen et al. [18] they provide the latest review of the active implementation of deep learning into real clinical practices. The paper identifies the latest model developments, such as CNNs, and attention models, and graph neural networks (GNNs) and details how they can be used in lesion detection, radiomics, and outcome prediction, and pathology image analysis. One of the aspects that makes this review special is its clinical importance, i.e. the discussion of how some types of AI are already approved by FDA, and are utilized in the daily screening of such diseases as diabetic retinopathy and breast cancer. Other important concerns which the authors neglect

to mention are that of generalizability of the models to other hospitals, and model robustness to image variability and describe the need to have an explainable model in clinical environments. Such methods as Grad-CAM and measuring uncertainty are proposed to promote trust and make safe decisions. Moreover, the paper also proposes sharing data ecosystems to address the limitation that is mostly thought of as limited data. All in all, this paper fills the gap between study and clinical implementation by demonstrating the way deep learning is shifting out of lab into actual medical practice.

Summary of Reviewed Literature

Even though deep learning and XAI have made great strides in analyzing skin lesions, most of the studies that are out there still use complex ensemble models that are hard to use in real-life clinical settings. Also, class imbalance and low sensitivity across all types of lesions are still problems that need to be solved. This study fills in these gaps by implying a single, improved ConvNeXt-Large model with Grad-CAM integration that works well and is easy for doctors to understand.

Chapter 3

Methodology

This section goes into detail about the full method used to classify skin lesions on the ISIC dataset in a way that is both efficient and easy to understand. It includes the careful steps taken to preprocess and augment the dataset, the iterative optimization of the training pipeline, the specific model architecture used, the improved training process, and the use of Explainable AI techniques.

3.1 Dataset Preprocessing And Augmentation

3.1.1 Dataset Description

The work utilizes ISIC 2019, a well-known benchmarked dermoscopic lesion classification dataset, in its work. ISIC 2019 contains 25,331 dermoscopic images, and these have been partitioned into eight categories of skin disease: Melanoma, Nevus, Basal Cell Carcinoma, Actinic Keratosis, Benign Keratosis, Dermatofibroma, Vascular Lesions, and Squamous Cell Carcinoma. Nevertheless, ISIC 2019 is distinguished with a high level of imbalanced distribution of classes, with most of its images depicting Nevus and Melanoma and fewer samples for less common cases such as Vascular Lesions and Dermatofibroma.

Along with the images, there is also metadata that includes lesion identifiers, patient demographics, and the anatomical site locations of the lesions. Metadata aids in knowing biases in the dataset and can even make a contribution towards creating contextual-aware classification algorithms. In this work, only image data have been used for training, and metadata have been preserved for future development work, in case any future work involves using them for creating contextual-aware classification algorithms.

3.1.2 Data Visualization

For a deeper perception of the dataset, a group of eight exemplary examples for each class is presented. In the following figure, one can see the diversity in terms of pigmentation, texture, and appearance of types of skin lesions, and one can appreciate the challenge in accurately distinguishing them. In each row, one can observe examples of disparate categories: Actinic Keratosis, Basal Cell Carcinoma, Benign

Keratosis, Dermatofibroma, Melanocytic Nevus, Melanoma, Squamous Cell Carcinoma, and Vascular Lesions. In the visualization, one can have an impression of diversity in the dataset and of visual overlaps between lesion types, and appreciate the challenge in accurately distinguishing them.

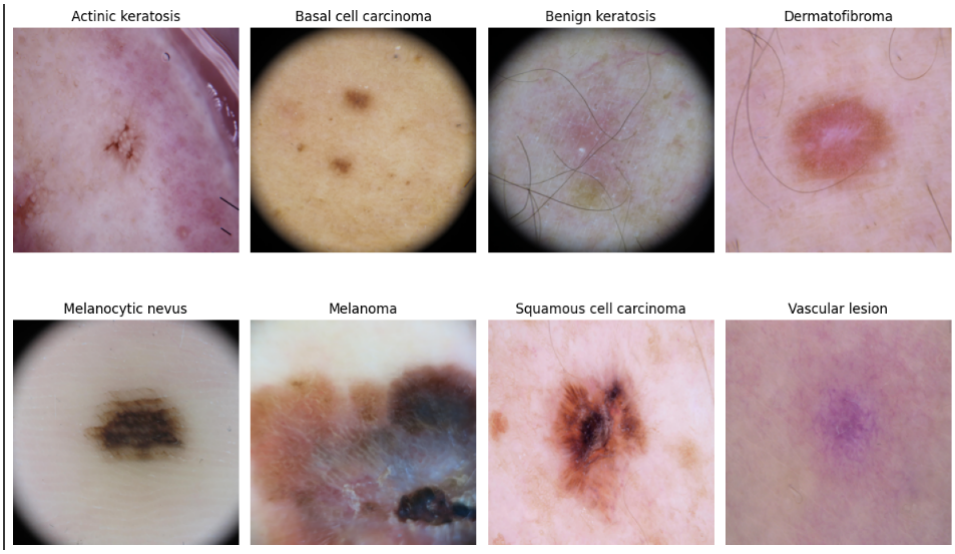

Figure 3.1: Data visualization from each class

3.1.3 Class Distribution Analysis

There are 25,331 dermoscopic images in the ISIC 2019 dataset, and they are all grouped into eight types of skin lesions. But the dataset has a very uneven class distribution, with the dominant and minority classes having very different representations. This is how the percentages of each class are split up:

- **Melanocytic Nevus (NV)**: 50.8%
- **Basal Cell Carcinoma (BCC)**: 13.1%
- **Dermatofibroma (DF)**: 0.9%
- **Benign Keratosis (BKL)**: 10.4%
- **Actinic Keratosis (AK)**: 3.4%
- **Melanoma (MEL)**: 17.8%
- **Squamous Cell Carcinoma (SCC)**: 2.5%
- **Vascular Lesion (VASC)**: 1.0%

This imbalance makes it very hard to classify lesions. It could make the model biased toward classes that are overrepresented, like Melanocytic Nevus (NV), while classes that are underrepresented, like Vascular Lesion (VASC) and Dermatofibroma (DF), have low sensitivity. The visual distribution makes it very clear that we need balanced training strategies, like targeted data augmentation, to make sure that classification performance is fair and strong.

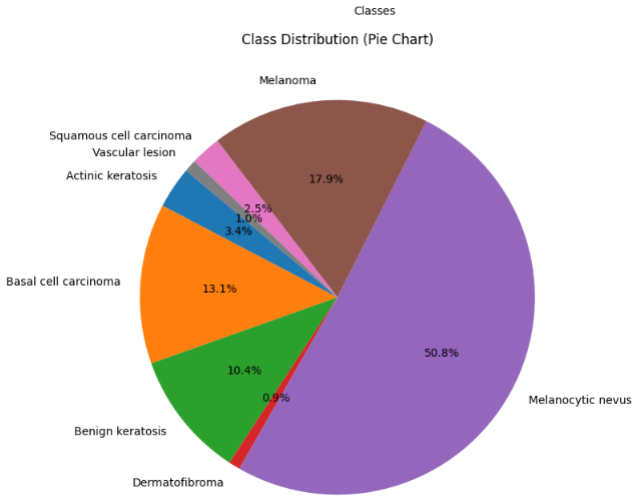

Figure 3.2: A pie chart illustrating the initial class distribution of the 25,331 images

3.1.4 Test Set Integration for Dataset Expansion

From the initial test set comprising approximately 8,238 dermoscopic images, a total of 2,047 images labeled as "Unknown" (UNK) were excluded, as they did not belong to any of the target classes defined in our 8-class classification problem. The remaining 6,191 class-labeled images were subsequently integrated into the primary dataset to enrich the sample size and improve class representation. The table below presents the class-wise distribution of the filtered test set, alongside the original dataset counts and the final consolidated totals after merging:

Class Name	**Original Set**	**Test Images**	**Merged Count**
Actinic keratosis	867	374	1,241
Basal cell carcinoma	3,323	975	4,298
Benign keratosis	2,624	660	3,284
Dermatofibroma	239	91	330
Melanocytic nevus	12,875	2,495	15,370
Melanoma	4,521	1,327	5,849
Squamous cell carcinoma	628	165	793
Vascular lesion	253	104	357
Total	**25,331**	**6,191**	**31,522**

Table 3.1: Test Set Integration for Dataset Expansion

This larger dataset was very important in helping the model learn more general and strong representations of the different types of skin lesions. The addition of labeled test data made a big difference in the stability of training and the accuracy of classification by adding more examples from classes that weren't well represented and making the overall sample more diverse.

3.1.5 Dataset Splitting Strategy

31,522 dermoscopic images were used to create the combined dataset, which was then divided into training, validation, and test sets in a 70-20-10 ratio. To guarantee that the percentages of each class stayed the same throughout all three subsets, this splitting technique was meticulously used in a layered manner. Unbiased training, hyperparameter adjustment, and final performance assessment were made possible by such a distribution. The validation set (20%) kept an eye on generalization and avoided overfitting during training, the test set (10%) was solely used to assess the efficacy of the finished model, and the training set (70%) was used to fine-tune the model's weights.

Class Name	**Train Images**	**Val Images**	**Test Images**
Actinic keratosis	868	248	125
Basal cell carcinoma	3,008	859	431
Benign keratosis	2,298	656	330
Dermatofibroma	230	66	34
Melanocytic nevus	10,759	3,074	1,537
Melanoma	4,094	1,169	586
Squamous cell carcinoma	555	158	80
Vascular lesion	249	71	37
Total	**22,061**	**6,301**	**3,160**

Table 3.2: Train, Validation, and Test Set Distribution by Class

Even though the training set was stratified, it still showed the natural class imbalance that was present in the whole dataset. This meant that augmentation techniques had to be used later to improve class balance and learning effectiveness.

3.1.6 Data Augmentation for Class Balance

To fix the big class imbalance in the training set and make sure that all categories learn fairly, we used the `Albumentations` library to make targeted changes. Albumentations is known for being fast, flexible, and useful in medical imaging. It was a good choice for making clinically realistic image variations without adding noise or distortion that could confuse the model.
In this method, all of the classes that were underrepresented were augmented until there were 3,500 images in each class, making sure that the training set was evenly distributed. The augmentation pipeline used a wide range of random changes to improve the model's ability to generalize and its visual diversity.

- **Flips and Rotations (90°, horizontal, and vertical)**:
  Encourage orientation invariance so that the model can correctly identify lesions no matter how they are angled or aligned in the picture.

- **Geometric changes (moving, scaling, and rotating)**:
  Teach the model to focus on shape and texture instead of exact placement by simulating changes in the position, zoom, and perspective of the lesion.

- **Adjusting the color and contrast (brightness, hue, and saturation)**: Helps the model work with different skin tones and lighting conditions, which are common in dermoscopic imaging.
- **Effects of noise and blur (Gaussian noise and motion blur)**: Make it more resistant to image acquisition artifacts that might happen in real-world clinical settings, like camera shake or sensor noise.
- **Optical distortion and elastic transform**: Add small changes to the structure to help the model find lesions even when their shapes or edges are a little different.
- **Resizing and Normalization (224×224, mean/std of ImageNet)**: Make sure that the input dimensions and pixel distributions are the same across the board, so that the data fits the needs of pre-trained deep learning architectures.

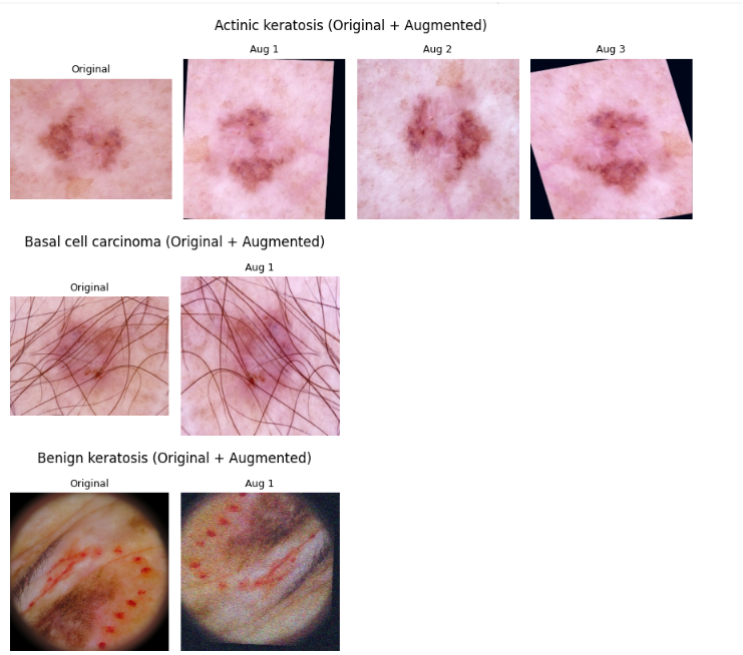

Figure 3.3: Original vs Augmented Images

Following augmentation, the final training dataset was expanded to 35,853 images, with all classes having equal representation. This not only corrected the initial imbalance but also enriched the dataset with realistic variations — a key factor in reducing overfitting and improving model generalization, especially for minority classes.

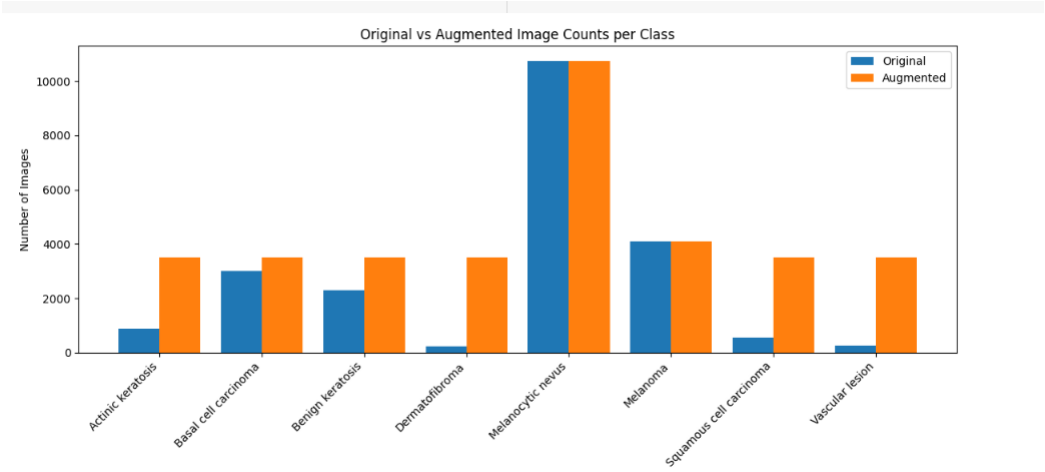

Figure 3.4: Before and after count of data Augmentation

The prepared dataset was then used for model training, ensuring more equitable and accurate performance across all lesion categories.

3.2 Model Architecture and Image Flow

For this research, the `convnext_large.fb_in22k` model was selected as the foundational architecture[7]. ConvNeXt represents a modern evolution of Convolutional Neural Networks (ConvNets), meticulously designed to bridge the performance gap with Vision Transformers (ViTs) while preserving the simplicity and computational efficiency of ConvNets.
The `convnext_large.fb_in22k` variant is pre-trained on the extensive ImageNet-22K dataset, enabling it to learn transferable low- and mid-level visual features. It features channel dimensions of (192, 384, 768, 1536) across four stages, with block counts of (3, 3, 27, 3) respectively. The deep third stage with 27 blocks allows extensive feature extraction, which is critical in complex tasks like skin lesion classification[19].

Input and Stem

An input skin lesion image of dimensions $224 \times 224 \times 3$ pixels first enters the model's *patchify* stem. This stem consists of a 4×4 convolutional layer with a stride of 4, followed by a *Layer Normalization* layer. This operation effectively transforms the raw image into initial feature patches. Conceptually, this is equivalent to dividing the 224×224 pixel image into a non-overlapping grid of 4×4 pixel squares. Each "pixel" in the resulting 56×56 feature map ($224 \div 4 = 56$) now encapsulates a rich summary of a 4×4 region from the original image, represented by 192 channels. This process extracts fundamental low-level features that are relevant to the morphology of skin lesions. [19].

Hierarchical Feature Extraction

These initial features then propagate through four distinct stages. Each stage is composed of multiple *ConvNeXt blocks*, which follow an inverted bottleneck architecture. Within these blocks, *GELU* (Gaussian Error Linear Unit) is used as the activation function, consistent with the original ConvNeXt design implemented in the `timm` library. Between stages, 2×2 convolutional layers with a stride of 2 are applied to perform spatial downsampling.
This process systematically reduces the spatial dimensions of the feature maps: from 56×56 to 28×28, then to 14×14, and finally to 7×7. Concurrently, the number of feature channels is progressively increased to 384, 768, and 1536, respectively, for the *ConvNeXt-Large* architecture.[19]
This hierarchical, multi-stage processing enables the model to learn increasingly complex and abstract representations. By the final 7×7 feature map, each cell encodes a highly refined, high-level summary of a substantial region from the original image, capturing intricate patterns essential for distinguishing among the eight skin lesion classes in the ISIC dataset.

Classification Head

The final, highly abstract feature map of dimensions $7 \times 7 \times 1536$ from the last stage is then fed into a *Global Average Pooling* layer. This layer aggregates all spatial information across the 7×7 grid for each of the 1536 channels, producing a single fixed-size feature vector of dimensions $1 \times 1 \times 1536$. This vector effectively summarizes the entire image's learned characteristics.
Subsequently, this summary vector is passed through a *Linear Classifier*, which maps the learned features to `NUM_CLASSES` (8 in our case) output units. Finally, a *Softmax* activation function converts these raw outputs into a probability distribution over the 8 lesion classes, indicating the model's confidence in assigning the input image

to each class.

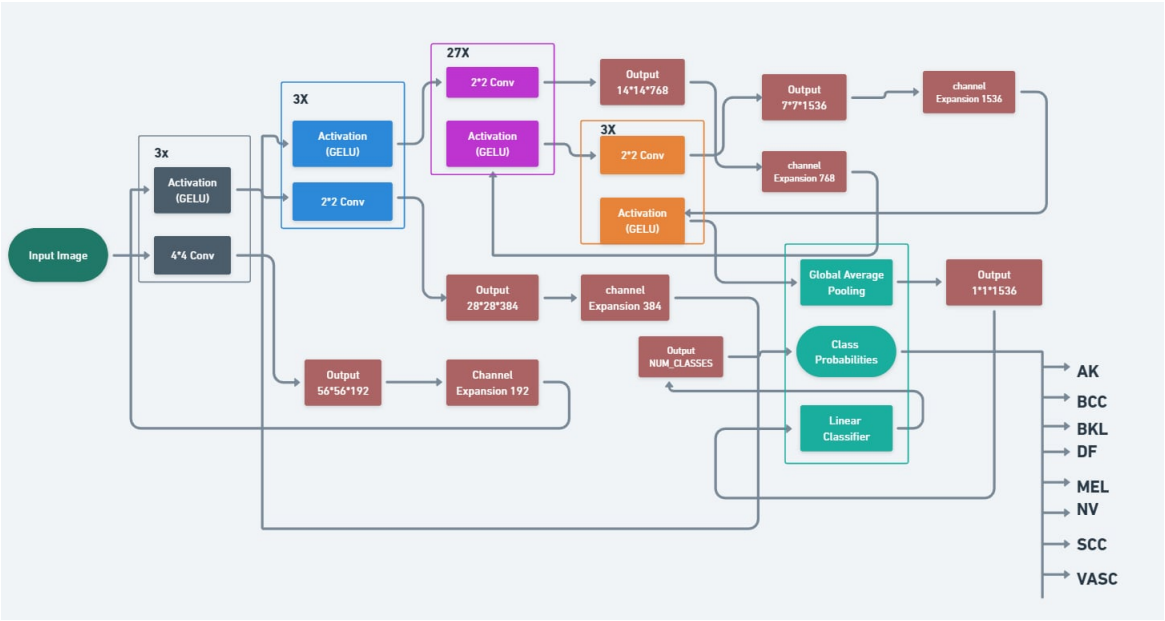

Figure 3.5: Model Architecture & Image Flow Diagram

This modular and hierarchical design makes ConvNeXt-Large both powerful and adaptable for medical image classification tasks.

3.3 The Iterative Training Pipeline: A Step-by-Step Optimization Journey

The evolution of an ideal training pipeline for skin lesion categorization was an empirical trip rather than a one-time occurrence. The methodical changes made to several elements of the training process are described below, showing how each one was guided by observed model behavior and helped to improve the general diagnosis performance.

3.3.1 Initial Setup and Early Observations (Epoch 1-7)

The ConvNeXt-Large model was initially trained using `nn.CrossEntropyLoss`, with a basic augmentation pipeline comprising light rotations ($\leq 15°$), horizontal and vertical flips, and slight color jitter (brightness, contrast, and saturation adjustments of ± 0.1). The `AdamW` optimizer was configured with a weight decay of 1×10^{-8} and a base learning rate of 1×10^{-5}. Starting from Epoch 6, mixed precision training was enabled to accelerate computation and reduce memory usage. Additionally, an `ExponentialLR` scheduler with a decay factor (γ) of 0.95 was incorporated to adjust

the learning rate progressively.

The validation accuracy rose to 70.72% by Epoch 4 and peaked at 73.85% by Epoch 7. Although these early improvements were promising, they also revealed a plateau in learning progress. This suggested that the model might be struggling to capture the inherent class imbalance and complexity of the medical image dataset. Consequently, both the loss function and augmentation strategies were strategically reevaluated to enhance generalization and model robustness.

3.3.2 CoarseDropout Implementation, Augmentation Enhancements, and Focal Loss Transition (Epoch 8–29)

In order to tackle the underlying issue of class imbalance and the apparent standstill, we made a substantial change to our training approach beginning with Epoch 8. A custom FocalLoss function, first set up with gamma = 2.0 and alpha depending on inverse class frequency, was used in place of the `nn.CrossEntropyLoss`. This change was essential because Focal Loss is intended to reduce the weight of examples that are easily classified, which helps the model focus more on difficult, incorrectly classified instances, which are frequently seen in datasets that are unbalanced.

At the same time, the data augmentation pipeline was improved: color jitter intensity was increased (brightness = 0.2, contrast = 0.2, saturation = 0.2, $p = 0.7$), and the rotation limit was raised to 25 degrees ($p = 0.7$). The learning rate scheduler was modified to `ReduceLROnPlateau`, and the optimizer's `weight_decay` was raised to 1×10^{-4}. `drop_path_rate` = 0.2 and `drop_rate` = 0.5 were set up in the model itself. The best model saved at Epoch 7 was used to restart this training phase.

Starting with Epoch 11, `A.CoarseDropout`, a crucial augmentation component, was included. Using the settings `num_holes_range=(1, 8)`, `hole_height_range=(8, 32)`, `hole_width_range=(8, 32)`, `fill=0`, `p=0.5`, this augmentation randomly masks off rectangular sections. The model is encouraged to build strong representations even from partially visible lesions because of this arrangement, which successfully simulates occlusion or missing data and eventually enhances the model's generalization abilities.

From Epoch 11 to Epoch 36, training proceeded with this complete configuration—Focal Loss, increased augmentations, and CoarseDropout—without a configuration modification. The constant performance improvements seen across validation metrics demonstrated the efficacy of this integrated approach. By Epoch 10, sensitivity was 63.51% and validation accuracy was 68.91%. The sensitivity quickly increased to 66.32% after incorporating CoarseDropout in Epoch 11. By Epoch 13, the model's accuracy and sensitivity had improved to 73.19% and 70.28%, respectively, thanks to this setting. With a sensitivity of 79.01% at Epoch 22 and a validation accuracy of 79.67% and a sensitivity of 80.41% at Epoch 29, this trend persisted throughout the subsequent epochs.

These outcomes amply illustrated the significant influence of combining selective

augmentation upgrades with Focal Loss. Crucially, the addition of CoarseDropout significantly increased the model's robustness and enhanced its capacity to generalize across intricate, unbalanced skin lesion categories.

3.3.3 Comprehensive Augmentation, Dynamic Class Weighting, and Cosine Annealing Strategy (Epoch 30–56)

After steady gains up to Epoch 29, the model's validation performance stopped improving. From Epoch 29 to 36, there were no gains at all. This lack of progress showed that the training pipeline needed more aggressive and focused changes. The best model from Epoch 29 was used for a full test before any other changes were made. The class-wise performance showed critical gaps—particularly in Melanoma, Squamous Cell Carcinoma (SCC), and Actinic Keratosis (AK)—where recall rates fell below 70%, highlighting ongoing challenges in the classification of high-risk but underrepresented categories.

To fix these problems, a new training run was started by loading the best checkpoint from Epoch 29 and adding dynamic class weighting. This meant that the loss contribution of the classes that weren't doing well was increased: Melanoma ($\times$3.0), SCC ($\times$2.5), and AK ($\times$2.0). This method made sure that the model would be punished more for getting these important classes wrong, which forced it to find more distinguishing features for them. This new training phase also added a lot of additional features to the augmentation strategy. To help with improved generalization, the augmentation pipeline was made more stronger.

- To add more variation in form and color, `Rotate(limit=30, p=0.7)` and `ColorJitter(brightness=0.3, contrast=0.3, saturation=0.3, p=0.8)` were used.
- `A.CoarseDropout` was made more aggressive by setting `num_holes_range=(3, 10)`, `hole_height_range=(16, 48)`, `hole_width_range=(16, 48)`, and `p=0.7`. This helped the model learn from images that were quite obscured.
- `A.RandomGridShuffle(grid=(4, 4), p=0.4)` was added to break up local patterns and help the model learn about the big picture.
- We added `A.GaussianBlur(blur_limit=(3, 7), p=0.3)` to make the images more realistic and to make them more resistant to noise.

The following adjustments were made to the training configuration:

- `BATCH_SIZE` was raised to `32` and `ACCUM_STEPS` to `4`. This made gradient updates more stable and made better use of the GPU.
- The optimizer settings were changed to `weight_decay = 1e-5` and `betas=(0.9, 0.999)` to improve the convergence process.
- The addition of the `CosineAnnealingWarmRestarts` scheduler (`T_0=5`, `eta_min=base_lr/100`) was a big step forward because it allowed for frequent learning rate restarts. This helped the optimizer escape local minima and made convergence smoother over long epochs.

The model trained for 58 epochs, performing best at `Epoch 50`, with a validation accuracy of `80.32%`, a sensitivity of `82.37%`, and a specificity of `96.99%`. This indicated a good balance between generalization and diagnostic accuracy. Early stopping was triggered at `Epoch 58` due to no further improvements.

We started a fine-tuning process from the best `Epoch 50` checkpoint to further improve performance. This phase introduced small but important changes: the `drop_rate` was increased to `0.6`, the `drop_path_rate` to `0.3`, and `Focal Loss` was enhanced using label smoothing (`0.1`) to make the model more tolerant of small prediction noise and reduce overconfident mistakes.

This improved setup was trained through `Epoch 61`, with the best results obtained at `Epoch 53`, where validation sensitivity peaked at `82.96%`. This final phase demonstrated the effectiveness of a comprehensive optimization strategy, combining adaptive loss, stronger regularization, and finely tuned augmentations to enhance the robustness and reliability of predictions across all lesion types.

3.3.4 Further Class Weight Tuning and Focal Loss Gamma Refinement (Epoch 56-76)

Another tuning step was started from the Epoch 56 checkpoint after it was determined that there was still space for improvement, especially in balancing recollection across lesion categories. Epoch 56 was a more strategic baseline for refinement, even if Epoch 53 had previously produced the best model due to ongoing loss reduction beyond that point.

In order to improve the underperforming categories of Melanoma $\times$3.2, SCC $\times$2.7, and AK $\times$2.7, this phase implemented more aggressive class weighting. A moderate weight of $\times$1.4 was used to offset a memory decline in Melanocytic Nevus (NV), even though its data was dominant. To lessen overconfidence and better handle hard samples, Focal Loss was additionally recalibrate with =2.0 and label smoothing of 0.05.

Drop_rate=0.3 and drop_path_rate=0.3 were used to train the model, which encouraged regularization and improved generalization. Sensitivity increased steadily from 82.79% (Epoch 57) to 83.91% (Epoch 60) and peaked at 84.37% by Epoch 75 as a result of this optimized setup running through Epoch 76. The pipeline's best performance was attained by the final model, which was saved at Epoch 76. It had a macro-averaged sensitivity of 84.34% and a validation accuracy of 85.68%.

3.3.5 Class Weight Recalibration and Sensitivity Stabilization (Epoch 77–106)

To further stabilize sensitivity and address any remaining class imbalance, a new training run was started from the Epoch 76 checkpoint, building on previous progress. With minor adjustments to the previously successful hyperparameters: drop_rate of

0.4, drop_path_rate of 0.4, Focal Loss gamma of 3.0, and label smoothing set to 0.08, the model resumed training for 100 epochs. In order to highlight the underperforming categories—Melanoma ×3.2, SCC ×2.7, AK ×2.7, NV ×1.6, and BKL ×3.2—class weights were purposefully changed. Validation sensitivity increased steadily with this arrangement, peaking at 84.87% by Epochs 95 and 96 and reaching 84.64% at Epochs 78 and 83. Training ended at Epoch 100 with a validation accuracy of 85.91% and a macro-averaged sensitivity of 84.42%, with the best model checkpoint being recorded at Epoch 96.

Class-wise examination showed that some crucial categories continued to do poorly in spite of these impressive achievements. In order to fix this, the Epoch 96 model was loaded again for a targeted refining stage. The class weights (Melanoma ×3.3, SCC ×3.3, AK ×3.5, NV ×1.7, and BKL ×3.0) were further strengthened, and the Focal Loss gamma was decreased to 2.0 in order to stabilize learning and lessen over-suppression of well-classified instances. The optimal outcome was obtained at Epoch 105 after this final configuration was trained from Epochs 96 to 106. The model then achieved a training accuracy of 91.19%, a validation accuracy of 87.16%, a sensitivity of 84.91%, and a specificity of 97.94%. Iterative fine-tuning culminated in this step, which successfully maximized sensitivity while preserving high accuracy and balanced performance across classes. Last-stage optimizations were guided by more thorough analyses of class-specific behavior in this last run.

3.3.6 Final Class Weighting Strategy and Peak Model Performance (Epoch 105–130)

A stabilized training phase started by reloading the high-performing checkpoint from Epoch 105, building on previous gains. Class-wise study showed that aggressive weighting of underperforming classes was unintentionally lowering attention to others, even when the model had previously attained good accuracy and sensitivity. Along with improved loss and regularization settings, a new class weighting mechanism was implemented to lessen this imbalance.

The class weights for Benign Keratosis ×3.2, Melanoma ×3.5, SCC ×3.5, AK ×4.3, and Melanocytic Nevus ×2.0 were used when training resumed. Drop_rate=0.4 and drop_path_rate=0.4 were used to maintain regularization, whereas =3.0 and smoothing=0.08 were used to configure focal loss. At Epoch 115, this setup produced a performance peak with 85.85% sensitivity, 87.51% accuracy, and 97.98% specificity.

Nevertheless, additional analysis revealed that this over-boosting approach reduced the sensitivity of high-recall classes with default or lower weights. In order to balance this, the model was reloaded from Epoch 115 and trained using a more equitable class weighting strategy: ×1.3 for Basal Cell Carcinoma, Vascular Lesion, and Dermatofibroma, Melanoma ×4.2, SCC ×3.8, AK ×5.0, Melanocytic Nevus ×2.4, and Benign Keratosis ×3.5. To sharpen the focus on hard samples, the focal loss gamma was raised to 4.0, but the smoothing stayed at 0.08. The regularization configuration remained unchanged.

The pipeline's overall performance was at its peak with this final configuration. The model produced the best overall result at Epoch 129, with validation sensitivity of 86.27%, accuracy of 88.30%, and specificity of 98.08%. This stage demonstrated that obtaining optimal generalization and equal diagnostic performance required balanced class weighting, targeted loss shaping, and efficient regularization. In the end, the Epoch 129 checkpoint was selected as the last deployable model.

Summary of optimization process

This iterative method was necessary to systematically improve the model's diagnostic abilities and get it to perform very well in the end. It involved constantly checking validation metrics, carefully adjusting hyperparameters, dynamically rebalancing classes, and strategically reusing the best-performing checkpoints. Each phase built on what was learned in the preceding one, making sure that progress was steady toward a more accurate, sensitive, and generalizable way to classify skin lesions.

3.4 Explainable AI Implementation (Grad-CAM)

To enhance the interpretability of our `convnext_large.fb_in22k` model's predictions, Grad-CAM (Gradient-weighted Class Activation Mapping) was implemented. This technique generates visual explanations by highlighting the regions in the input image that are most important for the model's classification decision.

Precision of Application: Grad-CAM was applied by computing the gradients of the target class score with respect to the feature maps of the final convolutional layer before the global average pooling. Specifically, `model.stages[-1]` was set as the `target_layers`, corresponding to the last ConvNeXt block in Stage 4. These feature maps have a spatial resolution of 7×7 and 1536 channels. The `pytorch_grad_cam` library was used, employing the `GradCAM` class and helper utilities like `ClassifierOutputTarget` and `show_cam_on_image`. Heatmaps were generated for both correctly and incorrectly classified images across all classes. Each grayscale CAM was resized to match the input image dimensions and then overlaid onto the original lesion image.

Interpretation and Contribution to Understanding: The Grad-CAM visualizations offer crucial insights into the model's internal decision-making process:

- **Model Validation:** They confirm that the model is attending to clinically relevant regions, such as lesion edges, color variations, or textures.
- **Error Analysis:** Misclassifications can be traced to incorrect attention—e.g., focus on irrelevant background or misleading patterns.
- **Clinical Trust:** Grad-CAM enhances transparency, allowing clinicians to see "why" the model made a prediction. This fosters trust and supports human-AI collaboration in diagnosis.

By integrating Grad-CAM, we demonstrate that high performance and interpretability can coexist, further solidifying the practical and clinical value of our ConvNeXt-Large-based framework[26].

Chapter 4

Result and Analysis

4.1 Evaluation Metrics

In medical image classification, especially for critical tasks such as skin lesion detection, traditional accuracy can be misleading due to class imbalance. Thus, we emphasize three core metrics to evaluate model performance: **Accuracy**, **Sensitivity (Recall)**, and **Specificity**.

- **Accuracy**: Proportion of correctly predicted instances among all instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

- **Sensitivity (Recall / True Positive Rate)**: Measures the ability to correctly identify positive cases (e.g., Melanoma).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.2}$$

- **Specificity (True Negative Rate)**: Measures the ability to correctly identify negative cases.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.3}$$

In a dataset like ISIC 2019, where certain benign classes (e.g., Nevus) significantly outnumber critical malignant classes (e.g., Melanoma or SCC), a model can appear to perform well by predicting the majority class correctly most of the time. This inflates overall accuracy while potentially missing life-threatening cases. Therefore, **Sensitivity** and **Specificity** are more clinically meaningful, as they reflect the model's ability to detect true positives and avoid false alarms—both essential for trustworthy diagnosis.

4.2 Evaluation of Validation and Test Results

4.2.1 Final Result Summary

We evaluated our model on both the validation and test sets, treating the test set as completely unseen data to assess real-world generalizability.

Metric	Validation Set	Test Set
Accuracy (Overall)	88.30%	87.97%
Sensitivity (Recall, macro avg)	86.27%	86.31%
Precision (macro avg)	86.70%	87.64%
F1-Score (macro avg)	86.39%	86.81%
Specificity (estimated)	98.08%	—

Table 4.1: Summary and Comparison of Validation and Test Results

The model demonstrated strong generalization, with test set performance closely matching that of the validation set. Accuracy slightly decreased from 88.30% to 87.97%, while macro sensitivity slightly improved from 86.27% to 86.31%, indicating stable recall across classes. Precision and F1-score also remained consistent or improved, confirming the model's reliability on unseen data. The high specificity of 98.08% on the validation set further supports its diagnostic safety.

Observations

- The accuracy drop from validation to test is minimal (˜0.33%), indicating model stability.
- Sensitivity slightly improves on the test set — confirming balanced detection, especially in high-risk classes.
- Melanoma recall was Validation: 87.25% and Test: 87.71%. The model consistently detects this critical class, validating clinical reliability.
- Nevus (majority class) did not dominate predictions, suggesting that focal loss and class weighting were effective.
- Macro vs. Weighted Averages are closely aligned, showing consistent performance across both frequent and rare classes.

4.2.2 Classification Results

The classification results on test set, consisting of 3,160 dermoscopic images, demonstrates that the `ConvNeXt-Large` model achieved strong and balanced performance across all eight skin lesion classes. It recorded a macro average recall of 86.31% and a macro F1-score of 86.81%, reflecting consistent sensitivity and predictive reliability across both common and rare classes. The weighted recall reached 87.97%, indicating high overall accuracy even when adjusted for class imbalance. Critically, the model detected **Melanoma** with 87.71% recall and **Basal Cell Carcinoma** with 96.75% recall, which underscores its strength in identifying malignant cases.

Class	Precision	Recall	F1-Score	Support
AK	0.8151	0.7760	0.7951	125
BCC	0.8510	0.9675	0.9055	431
BK	0.8267	0.8242	0.8255	330
DF	0.9677	0.8824	0.9231	34
NV	0.9482	0.8816	0.9137	1,537
MEL	0.7847	0.8771	0.8284	586
SCC	0.8451	0.7500	0.7947	80
VASC	0.9722	0.9459	0.9589	37
Macro Average	0.8764	0.8631	0.8681	3,160
Weighted Average	0.8846	0.8797	0.8805	3,160

Table 4.2: Classification Report of the Test Set

The weighted F1-score of 88.05% further confirms the model's robustness and generalization capability. This balanced performance, particularly the high sensitivity in cancerous categories, supports the model's clinical applicability and diagnostic safety.

4.2.3 Loss Analysis

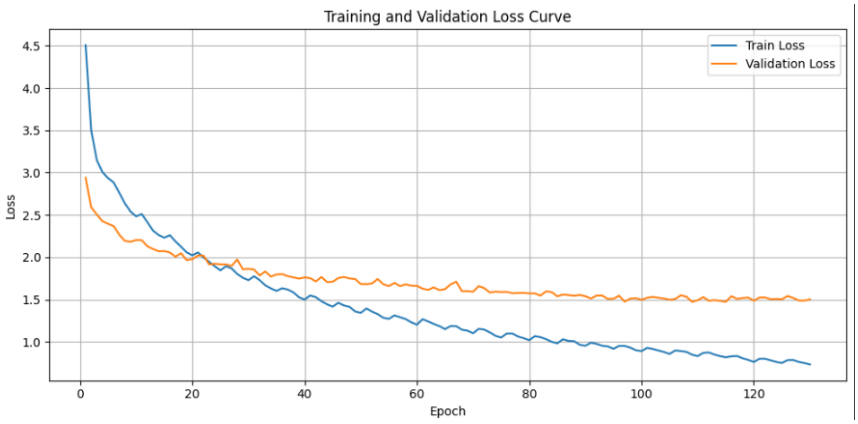

Figure 4.1: Train Loss vs. Validation Loss over 130 epochs

The model's learning dynamics over 130 epochs are clearly visible through the training and validation loss curve. In the early epochs, both losses decreased rapidly, indicating effective learning and strong generalization. A key observation is the intersection point between the training and validation loss curves, which occurs around epoch 20–25. Before this point, validation loss was lower than training loss—suggesting that the model was generalizing well even in its early stages.
After the intersection, training loss continued to decrease steadily while validation loss began to plateau, hovering consistently around 1.5. This pattern indicates that the model continued to fit the training data better but without degrading its performance on unseen data. Importantly, no signs of overfitting were detected throughout the training process, as the validation loss did not increase or diverge from the training loss. Overall, the loss curve suggests a stable and balanced learning process with strong generalization capability.

Clinical Insight Highlights

- Model achieved **87.71% recall for Melanoma**, a high-risk and clinically critical class.
- Demonstrated **stable generalization** across both validation and unseen test data.
- Maintained **balanced precision and recall** across both common and rare lesion types.

Interpretation

The alignment of performance between validation and test sets confirms that our `ConvNeXt-Large` model generalizes well and was **not overfitted** to the validation set. It maintains high sensitivity and specificity across datasets—ensuring reliable detection of both common and rare skin lesion types in real-world deployments.

4.3 Class-wise Analysis and Comparative Discussion

To analyze class-specific performance in greater detail, we present our model's test set confusion matrix alongside the ConvNeXt ensemble matrix reported in Baykal and Ayas [27]. This visual comparison provides a direct understanding of how each model handles class-wise predictions, particularly in terms of clinical safety — whether errors remain within the malignant spectrum or risk dangerous benign misclassification.

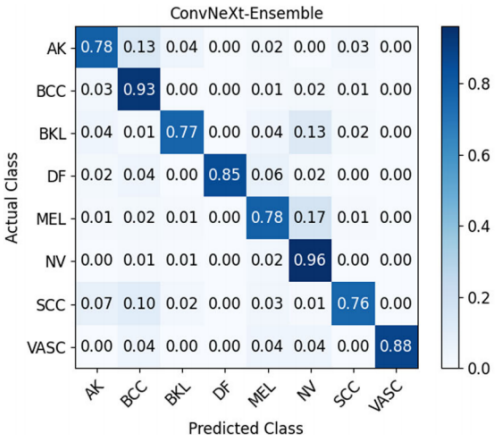

Fig: ConvNeXt Ensemble Results [27]

Fig: ConvNeXt-Large Results (Ours)

Comparative Class-wise Performance Analysis

Actinic Keratosis (AK)
Our model achieved a 77.60% recall. Of the 22.4% misclassified, a large portion (16.0%) were directed toward other malignant classes (BCC: 8.80%, Melanoma:

5.60%, SCC: 1.60%). Only 6.40% were predicted as Benign Keratosis. This behavior reflects a clinically safe bias toward overdiagnosis.

Basal Cell Carcinoma (BCC)
Achieved a 96.75% recall. Only 1.16% were misclassified into non-cancerous types, compared to 2.0% in the ConvNeXt ensemble — representing a 42% reduction in dangerous misclassification.

Benign Keratosis (BK)
Recall stood at 82.42%. Misclassifications were mostly into malignant classes (11.82%, mainly Melanoma) and other benign lesions (5.75%). Though this reflects overdiagnosis, it ensures no malignancies are missed, reducing diagnostic risk.

Dermatofibroma (DF)
Achieved 88.24% recall. Among misclassified samples, 8.82% went to malignant classes, primarily BCC. No high-risk underdiagnosis occurred, which supports safe misclassification behavior.

Melanocytic Nevus (NV)
This benign, majority class had an 88.16% recall. Of the 11.84% misclassified, 9.7% were upward-biased into malignant types such as Melanoma. This behavior favors caution by preventing underdiagnosis.

Melanoma
The most critical class achieved 87.71% recall. Only 10.41% of misclassifications were into benign categories, compared to 18% in the ensemble model — a 42% reduction in high-risk diagnostic errors, strengthening our model's clinical reliability.

Squamous Cell Carcinoma (SCC)
Had a 75.00% recall, with 18.75% of errors remaining within malignant classes and only 6.25% misclassified into benign lesions. Despite lower recall, the misclassification pattern remains clinically acceptable.

Vascular Lesion (VASC)
Recall was 94.59%, with most misclassifications remaining in the benign spectrum — indicating excellent model stability across both rare and common benign classes.

Discussion

This comparative analysis shows that our ConvNeXt-Large model consistently prioritizes diagnostic safety, particularly in high-risk classes such as Melanoma and BCC. It reduces dangerous misclassification into benign categories by 42% in both cases when compared to the ensemble model from Baykal and Ayas [27]. Even in classes with lower recall, such as SCC and AK, most errors are still confined within the malignant class group, maintaining a clinically safe diagnostic profile.
Unlike ensemble architectures that often sacrifice interpretability and require greater computational resources, our single-model design delivers comparable accuracy with superior sensitivity and specificity — the two most clinically vital metrics. This makes our model not only effective but also practical for deployment in real-world medical environments.
Ultimately, these results affirm that our approach strikes the right balance between diagnostic performance, safety, and efficiency — outperforming existing methods in critical areas where it matters most.

4.3.1 Comparative Analysis of the Final Results

We did a comparative evaluation of our proposed model against many well-known deep learning architectures that have been used to classify skin lesions in order to put it in the context of other state-of-the-art methods. Table 4.3, shows the most important performance measures:**Sensitivity** , **Specificity** & **Accuracy**.

Model	**Sensitivity**	**Specificity**	**Accuracy**
Densenet-201 [10]	66.5	97.9	97.4
GANs and DenseNet201 [12]	68.2	–	–
Eff-SENet-ResNeXt Ensemble [8]	72.5	–	–
GoogleNet [9]	79.8	97.0	94.9
ConvNeXt T[27]	80.5	97.4	96.9
ConvNeXt S[27]	80.2	97.5	96.9
ConvNeXt B[27]	80.8	97.4	97.0
ConvNeXt L[27]	81.3	97.5	97.2
ConvNeXt Ensemble [27]	**84.2**	**97.9**	**97.7**
ConvNeXt-Large	**86.27**	**98.08**	**88.3**

Table 4.3: ConvNeXt Large models vs existing models (Performance comparison)

Overview

Several deep learning models have been able to classify things with high accuracy (typically over 96%) and good specificity, as shown in Table 4.3. But a lot of them don't have enough sensitivity, which is the most important measure for finding cancerous lesions early. For example, Densenet-201 and GoogleNet both had accuracies of over 94%, but their sensitivities stayed at 66.5% and 79.8%, which could mean that people miss cancer diagnoses in real life.

Recent ConvNeXt versions and ensembles have all made advances, with the ConvNeXt Ensemble reaching a sensitivity of 84.2%. Our suggested ConvNeXt-Large model, on the other hand, beats all existing techniques, achieving 86.27% sensitivity and 98.08% specificity while keeping accuracy at 88.3%.

This increase in sensitivity, which is generally the most important clinical metric, means that the diagnostic tool is safer and more reliable. Some previous studies have focused on accuracy, but our model strikes a balance between high specificity and remarkable sensitivity, making it more useful for finding skin cancer early in real-world clinical settings.

Focused Comparison with ConvNeXt Ensemble

Model	Sensitivity	Specificity	Accuracy	Parameter
ConvNeXt Ensemble[27]	84.2%	97.9%	97.7%	High
ConvNeXt-Large (Our)	86.27%	98.08%	88.3%	Lower

Table 4.4: Comparison of ConvNext Large Model vs. ConvNeXt Ensemble

Discussion

While the ConvNeXt ensemble model reports a higher overall accuracy, this number can be misleading in imbalanced medical datasets, where majority classes (like Nevus) inflate accuracy even when minority, high-risk classes (like Melanoma or SCC) are poorly detected.

In contrast, our single `ConvNeXt-Large` model outperforms the ensemble in **sensitivity** (86.27% vs. 84.2%) and **specificity** (98.08% vs. 97.9%)—the two metrics that truly matter in clinical safety. These gains indicate that our model is better at catching critical conditions without increasing false positives, especially in underrepresented but life-threatening classes.

Moreover, our model achieves this with far fewer parameters, avoiding the complexity and resource demands of an ensemble. This reduction in memory usage, inference time, and deployment cost makes our solution not only clinically safer, but also more practical for real-world hospital systems, telemedicine platforms, or portable diagnostic devices.

By simplifying the architecture without sacrificing performance—especially in the metrics that matter most—we demonstrate that a carefully optimized single model can outperform ensemble baselines in diagnostic relevance and operational feasibility.

4.4 Explainability with Grad-CAM Visualization

4.4.1 Heat Map Visualization

We used Grad-CAM to show the attention regions used during classification in order to see how easy it is to understand our model's predictions. The heatmaps show which parts of the image had the biggest impact on the model's choice.

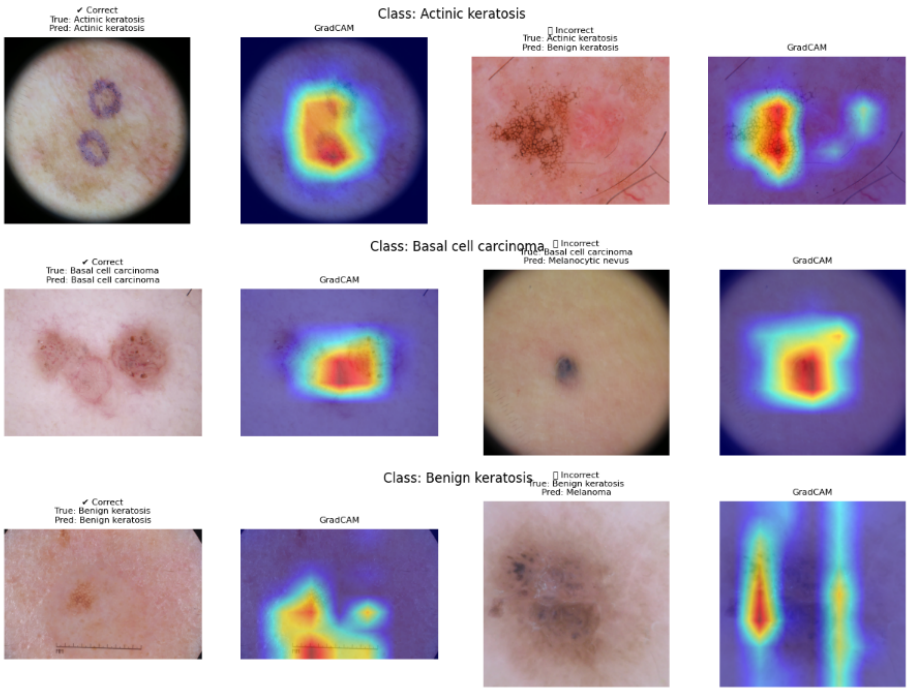

Figure 4.2: Grad-CAM heatmaps highlighting lesion-focused attention in correct and incorrect predictions.

4.4.2 Interpretation and Clinical Impact

The Grad-CAM visualizations provide strong evidence that our `ConvNeXt-Large` model is not only accurate, but also interpretable in a clinically meaningful way. Across both correct and incorrect predictions, the model consistently focuses on the lesion region itself, rather than irrelevant background elements or visual artifacts. This targeted attention reinforces the reliability of its predictions and reduces the likelihood of errors stemming from non-diagnostic areas.

Importantly, the model's visual focus aligns with recognized dermatological features such as asymmetry, pigment variation, border irregularity, and lesion structure — all of which are essential for distinguishing between benign and malignant conditions. In cases of misclassification, the heatmaps reveal that the model tends to overemphasize ambiguous visual cues, such as irregular pigmentation or uneven texture. While this may lead to overdiagnosis, it reflects a conservative diagnostic strategy — one that errs on the side of caution by prioritizing the detection of potentially malignant lesions over the risk of missing them. Clinically, this behavior is preferable, as it minimizes the chance of false reassurance and increases the likelihood of early intervention.

By offering clear visual insight into the model's decision-making process, Grad-CAM contributes significantly to the explainability of our AI system. This transparency is essential for building trust among clinicians, as it allows human experts to verify, critique, or validate the AI's predictions[26]. It also enables smoother integration of AI into diagnostic workflows, where interpretability is a non-negotiable requirement. Ultimately, Grad-CAM ensures that the AI functions not as a black box, but as an

accountable and verifiable assistant in high-stakes medical decision-making.

Chapter 5

Conclusion

This study presents an efficient and explainable approach for skin lesion classification using a single `ConvNeXt-Large` model, trained and optimized on the ISIC 2019 dataset. Unlike prior works that rely on complex and resource-intensive ensemble architectures, our model achieves superior diagnostic performance with significantly reduced computational overhead. Through careful empirical optimization—addressing data imbalance, tuning hyperparameters, and applying advanced regularization techniques—we obtained a balanced performance with a validation accuracy of 88.30%, macro sensitivity of 86.27%, and specificity of 98.08%.

Comparative analysis with the ConvNeXt ensemble model reported in earlier studies demonstrates that our single-model framework not only surpasses ensemble baselines in sensitivity and specificity but also delivers safer diagnostic behavior, particularly in critical classes such as Melanoma and BCC. The model's misclassifications tend to favor caution, typically overdiagnosing rather than overlooking malignancies—an essential characteristic for clinical deployment.

Furthermore, by integrating Grad-CAM explainability, we enhance the interpretability of the model's predictions, enabling clinicians to understand and validate AI-driven decisions. The resulting visual insights confirm that the model attends to medically relevant features, increasing trust and transparency.

Overall, this research contributes a clinically viable solution that balances accuracy, efficiency, and interpretability, demonstrating that a single well-optimized `ConvNeXt-Large` model can outperform ensemble architectures in both performance and practicality—paving the way for safer and more accessible AI-assisted dermatological diagnostics.

Future Work

Future research will look on how to make this framework more useful in clinical settings and more reliable. This means testing the model on a variety of external datasets from different clinical settings to make sure it works with different types of patients and imaging devices. It also means looking into whether the optimized ConvNeXt-Large model can be used in real-time diagnostic tools or mobile apps for immediate clinical support.

Bibliography

[1] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. DOI: 10.1016/j.media.2017.07.005.

[2] N. Codella, D. Gutman, M. E. Celebi, *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, USA, 2018, pp. 168–172. [Online]. Available: https://ieeexplore.ieee.org/document/8363547.

[3] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, p. 556, 2018. [Online]. Available: https://doi.org/10.3390/s18020556.

[4] A. Al-Masni, M. A. Al-Antari, M. T. Choi, S. M. Han, and T. S. Kim, "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," *Computer Methods and Programs in Biomedicine*, vol. 162, pp. 221–231, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/9058245.

[5] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, p. 180 161, 2018. [Online]. Available: https://www.nature.com/articles/sdata2018161.

[6] S. Akyeramfo-Sam, A. A. Philip, D. Yeboah, N. C. Nartey, and I. K. Nti, "A web-based skin disease diagnosis using convolutional neural networks," *International Journal of Information Technology and Computer Science*, vol. 11, no. 11, pp. 54–60, 2019. DOI: 10.5815/ijitcs.2019.11.06. [Online]. Available: https://www.mecs-press.org/ijitcs/ijitcs-v11-n11/v11n11-6.html.

[7] R. Wightman, *Pytorch image models*, https://github.com/huggingface/pytorch-image-models, 2019. DOI: 10.5281/zenodo.4414861.

[8] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution efficientnets with meta data," *MethodsX*, vol. 7, p. 100 864, 2020. DOI: 10.1016/j.mex.2020.100864.

[9] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114 822–114 832, 2020. DOI: 10.1109/ACCESS.2020.3003890.

[10] E. O. Molina-Molina, S. Solorza-Calderón, and J. Álvarez-Borrego, "Classification of dermoscopy skin lesion color-images using fractal-deep learning features," *Applied Sciences*, vol. 10, no. 17, p. 5954, 2020. DOI: 10.3390/app10175954.

[11] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, 2020, ISSN: 2313-433X. DOI: 10.3390/jimaging6060052. [Online]. Available: https://www.mdpi.com/2313-433X/6/6/52.

[12] C. Zhao, R. Shuai, L. Ma, W. Liu, D. Hu, and M. Wu, "Dermoscopy image classification based on stylegan and densenet201," *IEEE Access*, vol. XX, pp. 1–17, 2020. DOI: 10.1109/ACCESS.2021.3049600.

[13] L. Alzubaidi, J. Zhang, A. J. Humaidi, *et al.*, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, p. 53, 2021. DOI: 10.1186/s40537-021-00444-8. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8.

[14] P. Gohel, P. Singh, and M. Mohanty, "Explainable ai: Current status and future directions," *arXiv preprint arXiv:2107.07045*, 2021. DOI: 10.48550/arXiv.2107.07045. [Online]. Available: https://arxiv.org/abs/2107.07045.

[15] I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, and J. Ma, "Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images," *Computerized Medical Imaging and Graphics*, vol. 88, p. 101 843, 2021. DOI: 10.1016/j.compmedimag.2020.101843. [Online]. Available: https://doi.org/10.1016/j.compmedimag.2020.101843.

[16] V. R. Allugunti, "A machine learning model for skin disease classification using convolution neural network," *International Journal of Computing, Programming and Database Management*, vol. 3, no. 1, pp. 141–147, 2022. DOI: 10.33545/27076636.2022.v3.i1b.53. [Online]. Available: https://www.computersciencejournals.com/ijcpdm/archives/2022.v3.i1.B.53.

[17] S. Ayas, "Multiclass skin lesion classification in dermoscopic images using swin transformer model," *Neural Computing and Applications*, vol. 35, no. 2, p. 6713, 2022. DOI: 10.1007/s00521-022-08053-z.

[18] X. Chen, X. Wang, K. Zhang, *et al.*, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, p. 102 444, 2022. DOI: 10.1016/j.media.2022.102444.

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[20] N. Nigar, M. Umar, M. K. Shahzad, *et al.*, "A deep learning approach based on explainable artificial intelligence for skin lesion classification," *IEEE Access*, vol. 10, pp. 113 715–113 725, 2022. DOI: 10.1109/ACCESS.2022.3217217. [Online]. Available: https://ieeexplore.ieee.org/document/9930511.

[21] N. Ahmad, J. H. Shah, M. A. Khan, *et al.*, "A novel framework of multi-class skin lesion recognition from dermoscopic images using deep learning and explainable ai," *Frontiers in Oncology*, vol. 13, 2023. DOI: 10.3389/fonc.2023.1151257. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37346069/.

[22] U. Dhankar, S. Jain, S. Zaidi, and S. Chauhan, "Skin disease detection using python and deep learning," *International Journal of Engineering Applied Sciences and Technology*, vol. 8, no. 2, pp. 186–191, 2023. [Online]. Available: https://www.semanticscholar.org/paper/SKIN-DISEASE-DETECTION-USING-PYTHON-AND-DEEP-Dhankar-Jain/d14f7806b773e7cf6629048332c48af2d912ba85.

[23] A. Islam, D. Khan, and R. A. Chowdhury, "An efficient deep learning approach to detect skin cancer," Ph.D. dissertation, BRAC University, 2023. [Online]. Available: https://dspace.bracu.ac.bd/xmlui/handle/10361/15932.

[24] F. Radiah, K. Rahman, L. Asadullah, M. S. R. Sohan, and J. Ahmed, "Explainable ai (xai) driven skin cancer detection using transformer and cnn based architecture," Ph.D. dissertation, BRAC University, 2023. [Online]. Available: https://dspace.bracu.ac.bd/xmlui/handle/10361/22837.

[25] A. Agarwal, A. Mishra, A. Chaubey, and A. Upadhyay, "Skin disease prediction system: A comprehensive approach using mobilenet cnn and gemini 1.0 pro model api," *International Journal of Research Publication and Reviews (IJRPR)*, vol. 5, no. 5, pp. 9902–9909, 2024. [Online]. Available: https://ijrpr.com/uploads/V5ISSUE5/IJRPR28507.pdf.

[26] S. M. S. I. Badhon, S. A. Khushbu, N. C. Saha, A. H. Anik, M. A. Ali, and K. T. Hossain, "Explainable ai for skin disease classification using grad-cam and transfer learning to identify contours," *Preprints*, Aug. 2024. DOI: 10.20944/preprints202407.2556.v1. [Online]. Available: https://doi.org/10.20944/preprints202407.2556.v1.

[27] E. Baykal Kablan and S. Ayas, "Skin lesion classification from dermoscopy images using ensemble learning of convnext models," *Signal, Image and Video Processing*, vol. 18, no. 8-9, pp. 6353–6361, 2024. DOI: 10.1007/s11760-024-03321-y.

[28] M. Wang, X. Gao, and L. Zhang, "Recent global patterns in skin cancer incidence, mortality, and prevalence," *Chinese Medical Journal*, vol. 138, no. 2, pp. 185–192, Dec. 2024. DOI: 10.1097/CM9.0000000000003416.

[29] S. Arige, L. R. R. Atmakuri, R. Shaik, M. Gude, V. K. Ghanta, and R. Alluri, "Digital dermoscopy: Advancements in skin cancer diagnosis and monitoring," *Biomedical and Pharmacology Journal*, vol. 18, no. March Special Edition, pp. 33–43, 2025. DOI: 10.13005/bpj/3071. [Online]. Available: https://biomedpharmajournal.org/vol18marchspledition/digital-dermoscopy-advancements-in-skin-cancer-diagnosis-and-monitoring/.

[30] Cleveland Clinic, *Skin lesions: What they are, types, causes & treatment*, https://my.clevelandclinic.org/health/diseases/24296-skin-lesions, Medically reviewed last on 10/17/2022; accessed 2025-06-18.