2026-08

Scalp disease classification using modern and traditional deep learning architecture

Mobasher, Gazi Muhammad

Independent University, Bangladesh (IUB)

SCALP DISEASE CLASSIFICATION USING MODERN AND TRADITIONAL DEEP LEARNING ARCHITECTURES

August 2026

Prepared by:

Gazi Muhammad Mobasher
ID: 2221407

Mohammed Raihan Ur Rashid
ID: 2030571

Asif Karim Alif
ID: 2220031

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by:

Saadia Binte Alam
Associate Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

A Final Year Design Project (FYDP) Report submitted for the Degree of Bachelor's of Computer Science & Engineering

Attestation

We know that any form of plagiarism is not permitted and conflicts with the academic policies of our university. We attest that the work presented is original and that it was created by us. This project has some copyrighted materials, publicly available datasets, pre-trained models, and external references, all of which have been cited in accordance with accepted international citation guidelines.

Author Name:

.....

Signature:

.....

Author Name:

.....

Signature:

.....

Author Name:

.....

Signature:

.....

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

We are most grateful to our head supervisor, Dr. Saadia Binte Alam, Associate Professor. for providing us with continuous supervision in the phases of our Final Year Design Project, and we are also grateful to Md. Rashedur Rahman, D.Eng, and research instructors Samiul Karim Mazumder, and Siam Tahsin Bhuiyan, for their continuous guidance, comments, and support throughout this research work. The suggestions, recommendations, and time provided during the development of the study played an indispensable role in shaping the final project report.

We also wish to thank the Department of Computer Science and Engineering at Independent University, Bangladesh, for providing an academic environment where this research could be carried out, and the CCDS (Center for Computational and Data Sciences), for their exposure to research materials, computation resources and rigorous academic feedback significantly contributed to the completion of this project.

Moreover, we would like to acknowledge the researchers and contributors whose publicly available scalp and hair image dataset made this benchmark study possible. We also acknowledge the developers of PyTorch and the open-source deep learning community, whose tools and model implementations supported the experimental work.

Finally, we used AI-powered language tools, including ChatGPT, only to polish phrasing and improve the quality of this thesis report. These tools were not used to build the dataset, evaluate or formulate any experimental metrics, images, or the scientific analysis reported in this work.

List of Publications

G. M. Mobasher, M. R. U. Rashid, A. K. Alif, S. K. Mazumder, R. Islam, R. Rahman, A. Islam, and S. B. Alam, "Scalp Disease Classification Using Modern Deep Learning Architectures: A Benchmark Study," 2026, International Conference on Engineering and Frontier Technologies, (ICEFronT), Tangail, Bangladesh, Jun. 25-26, 2026. (Accepted)

Abstract

Scalp diseases are dermatological conditions that affect both physiological functioning and psychological health. Adept classification of these conditions can support early diagnosis, reduce diagnostic delay, and assist patients in making more pre-emptive decisions from their diagnosed scalp images. Traditional diagnosis is often dependent on expert inspection and can be delayed due to the analysis of overlapped symptoms, subtle visual patterns, and limited availability of dermatology specialists. Therefore, the classification of deep learning-based scalp diseases has become an important research direction in medical vision.

This thesis report presents a comprehensive benchmark study of seven modern deep learning architectures for six-class scalp disease classification. The evaluated models consisted of six CNNs and one Transformer (DenseNet121, EfficientNetB0, EfficientNetB2, ConvNeXt-B, ResNet34, ResNet50, ViT-B/16). The dataset was obtained from Roboflow Universe's Hair-Scalp-Analysis v2 scalp and hair image dataset and refined into six classes: Healthy Scalp, Alopecia, Folliculitis, Dermatitis, Dandruff, and Hair Loss. The final dataset contained 1368 images and was split into training, validation, and testing subsets in a 70%, 15%, and 15% ratio, respectively. Each image was resized to 224 x 224 pixels before model training. Additionally, an augmented version of the dataset was also created to address the class imbalance within the original dataset. After applying various image transformations such as random rotation, random zoom or crop, vertical flip, horizontal flip, affine transformation, brightness adjustment, and contrast adjustment, the total number of images in the augmented dataset was 2,347.

The performance of each model was evaluated using Accuracy, Precision, Recall, Specificity, F1-score, and AUROC. For the dataset without augmentation, the findings show that ConvNeXt-B was the most balanced model, achieving 92.23% of test accuracy, along with the following macro-averaged metrics: 89.10% precision, 86.39% recall, 92.20% specificity, and the highest F1-score of 84.90%. EfficientNetB2 achieved the same accuracy of 92.23% and the highest macro-averaged specificity of 96.10%, while ResNet34 obtained the highest macro-averaged AUROC value of 0.9785. Further experimentation with the augmented dataset showed improved classification capability, with EfficientNetB2 achieving the highest accuracy of 93.69%, an AUROC of 0.9873, an F1-score of 88.40%, and a specificity of 98.70%. ConvNeXt-B followed closely with an accuracy of 92.23%. These results show the effects of augmentation on a dataset.

Overall, the results suggest that convolutional neural networks, particularly modern CNN designs with transformer-inspired functions, provide strong credibility for scalp and hair disorder classification, especially when the dataset is relatively small and visual details are significantly important. Furthermore, the study emphasizes the importance of data augmentation for increasing model performance in the case of small medical image datasets that have class imbalance.

Contents

List of Figures

List of Tables

Chapter-1: Introduction

1.1 Background

Scalp is the skin at the top of the human head, which is usually higher in density of follicles. The hair density in this part of the body creates warm, dark and humid conditions on the surface of the scalp which produces the optimum environment for microbes to grow and multiply. In addition, the adult scalp produces more sebum, which can lead to a significant amount of bacteria and fungus. The scalp also regularly comes in contact with external objects while brushing or other styling purposes causing friction injuries and introducing microorganisms. Therefore, the scalp is vulnerable to various mycotic diseases including dandruff, seborrheic dermatitis and tinea capitis, parasitic infestation (pediculosis capitis) and inflammatory diseases (psoriasis) [1]. Scalp and hair disorders are dermatological conditions characterized by a variety of symptoms such as itching, scaling, redness, follicular inflammation, and hair loss [2]. The problems are caused by overgrowth of microorganisms like the fungus Malassezia, which causes dandruff, or the bacterium Staphylococcus, which causes unregulated inflammation and folliculitis. Scalp and hair disorders are a growing health concern worldwide, thought to be caused by modern lifestyle challenges such as chronic stress, dietary habits, and increased exposure to environmental pollutants [2]. One of the difficulties in their diagnosis is that many diseases have very similar visual symptoms, and it is hard to classify them without expert clinical evaluation [3]. The scalp is a complex structure, and the hair is dense, making it difficult to detect and identify subtle signs of disease during a routine examination [3] [4].

Appearance of hair and scalp is linked to identity, and conditions such as alopecia and dandruff can diminish self-confidence and social interaction [5][6]. But early detection is not only important for the treatment of surface symptoms, but also because some scalp conditions may indicate underlying health problems, such as autoimmune disorders or nutritional deficiencies, that warrant immediate medical attention [5]. Traditional diagnosis is based on visual inspection, dermoscopic analysis, and clinical judgment [3]. This may be a lengthy and unreliable process, especially if several conditions result in similar visual symptoms [3]. An automated classification system could provide dermatologists with a reliable second opinion and alleviate the burden of preliminary screening in a meaningful way [3][4]. This is particularly well suited for deep learning as it can learn useful visual features directly from the raw image data, without relying on hand-engineered descriptors [7][8].

Scalp and hair disorders belong to the category of the most frequent dermatological disorders that affect individuals of different ages and populations all over the world [9].

The World Health Organization (WHO) approximated that about 70% of adults all over the world experience some form of scalp or hair problem [9]. Among those, Alopecia has the most significant impact on people worldwide, affecting millions across various demographics. In South Korea, the registered population of alopecia patients nearly doubled from 103,000 in 2001 to 233,000 in 2020, with the true domestic figure estimated at approximately 10 million, considering unaccounted cases, stemming from age-related and genetic factors [10]. An autoimmune variant of this condition, alopecia areata, affects approximately 700,000 people in the United States, while another variant of the condition, androgenic alopecia, impacts roughly 40% of women and 70% of men worldwide [11]. Besides alopecia, dandruff affects a significant portion of the people worldwide as well, affecting about 50% of the global adult population [12]. Studies have also shown that dandruff also affects around 18% of children in the United States and Australia, revealing that scalp and hair disorders impact everyone regardless of their age [9].The conditions, which include alopecia, dermatitis, folliculitis, and dandruff, usually show symptoms such as the loss of hair, inflammation, scaling, redness, and irritation [2]. These disorders are not life-threatening, but they can greatly influence the quality of life of a particular person and, in some cases, may be the manifestation of other conditions, including autoimmune diseases and other systemic abnormalities [5]. Accordingly, timely diagnosis is significant in providing timely treatment to patients, averting the progression of diseases, and enhancing patient outcomes [5].

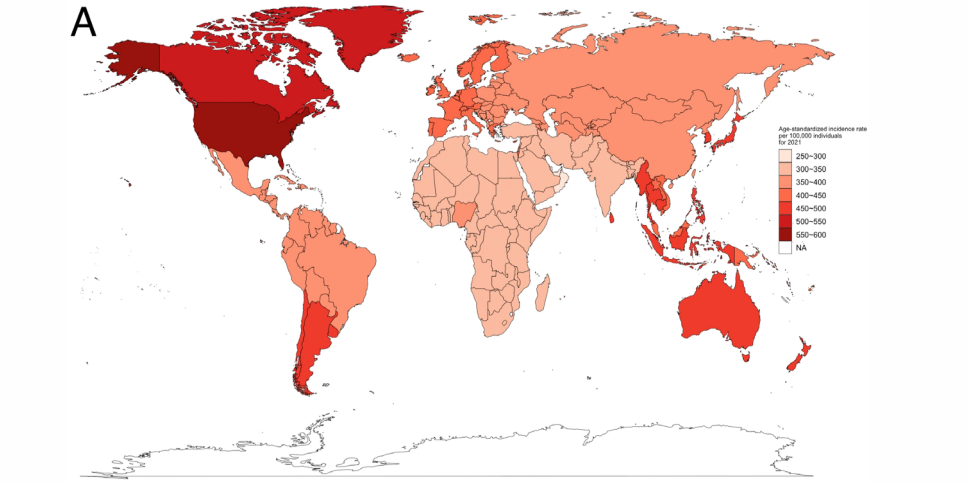

Figure 1: Illustration A shows the age-standardized incidence rate of Alopecia Areata per 100,000 individuals based on a 2021 study [6].

Scalp disorders are mainly diagnosed through visual examination, dermoscopic evaluation, and clinical prowess [3]. A number of scalp conditions, however, have a similar appearance, and distinctions between them are challenging even to experienced practitioners [3]. The differences in lighting conditions, hair density, disease severity, and image quality add to the complexity of diagnosis further [3]. Consequently, manual evaluation may be time-consuming and may vary between observers [3]. These issues have prompted the creation of computer-aided diagnostic systems that can help clinicians in the screening and diagnostic process [3].

Convolutional Neural Networks (CNNs) have shown excellent results in medical image analysis, capturing spatial patterns at various levels [8][13]. The initial layers of the CNN extract simple edges and textures, and the subsequent layers extract more complex

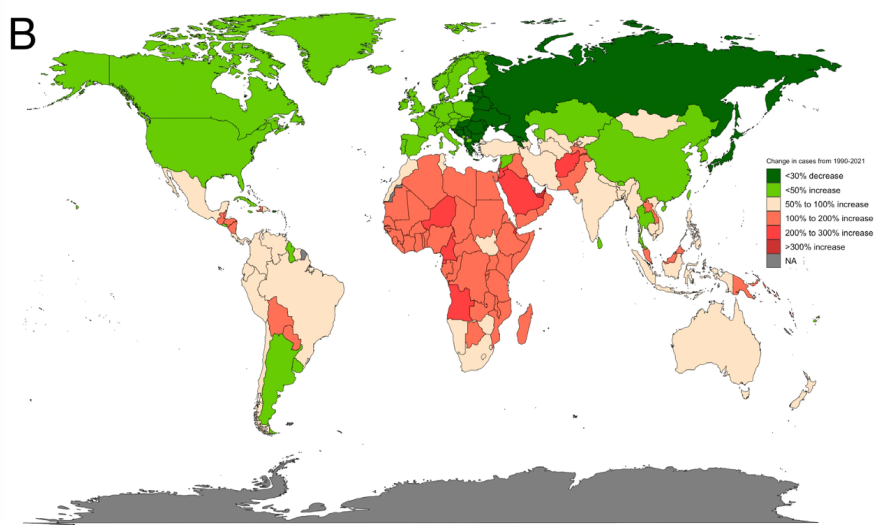

Figure 2: Illustration B shows the trend in cases of Alopecia Areata from 1990 to 2021 based on a 2021 study [6].

structural features [8]. These are useful for the classification of scalp diseases since diagnostic clues may manifest as focal areas, scaling, lesions, redness, alteration of hair density, or follicular texture [7]. Newer architectures like EfficientNet and ConvNeXt-B are based on classical CNN architectures, but introduce efficient scaling strategies and design principles based on recent research [14][15]. Recently, transformer-based vision architectures have also been a focus of attention in medical imaging [16]. Vision Transformers break up an image into fixed patches and use attention mechanisms to capture relationships between far away regions of the image [16]. However, these models require a large amount of data to learn stable representations, which can be a problem in the case of smaller medical datasets [16]. This is worth comparing the model of the transformer with CNN based models for the classification of scalp diseases [2][7].

The most recent breakthroughs in computer vision and deep learning have revolutionized the process of medical diagnosis by image [8]. Deep learning models have the ability to learn hierarchical visual representations automatically on raw image data, eliminating the necessity of manually designed feature extraction algorithms [8]. CNNs are among these methods and have shown success in a variety of medical imaging applications, such as disease detection on radiological and clinical images [13]. This is why they are skilled at locating hidden spatial intricacies within scalp images, where features for diagnosis exist as localized textures, lesions, inflammation, scaling patterns, or variations in hair density [3]. CNN-based methods have been proven to be relevant in the previous studies to automatically identify the scalp diseases with high classification accuracy across different scalp diseases [2].

The advent of more recent deep learning models has increased the number of approaches that can be used to classify images. DenseNet enhances information exchange by dense feature reuse between network layers, and EfficientNet proposes a compound scaling scheme, balancing network depth, width, and image resolution to achieve better predictive performance with computational efficiency [17][14]. More recently, ConvNeXt-B has updated the traditional CNN architecture with an architecture driven by transformer-based models [15]. Meanwhile, Vision Transformers (ViTs) have launched self-attention mechanisms that are able to capture long-range relationships between image regions and

have shown competitive performance in medical image analysis tasks [2][16]. The recent increase in the variety of deep learning architectures has provided new potential to advance automated classification of scalp diseases and should be further explored regarding their performance relative to each other in the field.

1.2 Motivation

The need for a reliable and reproducible bench mark model for the classification of scalp diseases is the motivation behind this research. Despite its wide use in many fields of medical image analysis, the problem of scalp and hair disorder classification is still a relatively under-explored field compared to other more well-established problems like chest X-ray analysis, skin lesion classification and organ segmentation.

The main contribution and driving force behind this work is the challenge of diagnosis of scalp diseases. Many different disorders may have similar visual presentation and the image quality or hair covering may further complicate the evaluation. However, a computational system trained with visual patterns from well-designed databases may alleviate this uncertainty and facilitate earlier and more regular screening.

Secondly, deployment is one of the important motives of the study. For a real time scalp disease classification system, one should consider a balance of predictive capacity, robustness, and computing efficiency. Depending on the specific use case, lighter models like EfficientNetB0 and EfficientNetB2 could be more suitable for use in mobile or resource-limited settings, whereas models like ConvNeXt-B might be better suited for more resource-rich environments that demand higher accuracy and more extensive feature representation.

1.3 Research Questions

The above motivation motivates the following research questions. Based on the above, the following research questions are formulated. Instead of just reviewing a single architecture, this research asks the questions under a common experimental setup in which the models chosen are tested under uniform dataset, pre-processing, training, and evaluation conditions.

1. **RQ1:** How does modern and traditional deep learning architectures classify six scalp condition classes under a standard experimental setup.
2. **RQ2:** What is the impact of data augmentation on the classification accuracy of the chosen deep learning architectures in the classification of scalp diseases?
3. **RQ3:** Under which performance metrics (Accuracy, Precision, Recall, Specificity, F1-score, and AUROC), CNN architectures and Vision Transformers approach each other, and in which do they diverge, especially in the case of class imbalance and visually similar scalp conditions?

RQ1 aims at providing a baseline comparison of the performance of seven deep learning architectures using a same experimental setup. By making the comparisons, the influence of architectural design can be studied, but differences in training conditions are minimized.

RQ2 explicitly investigates the effect of data augmentation on the problem of data imbalance on the current dataset. The augmentation can be used to add more diversity to the training examples when the images of the scalps exhibit variations in orientation, appearance, and other visual characteristics. This comparison between baseline and augmented experiments enables the changes due to augmentation to be quantified, rather than assuming that augmentation always results in an improvement.

RQ3 addresses the need for a comprehensive evaluation of classification performance. In cases where class distributions are not equal, or specific classes are harder to discriminate, accuracy may not suffice as a representation of model behavior. For this reason, in addition to Accuracy, Precision, Recall, Specificity and F1-score are also taken into account, as well as the AUROC index which is a measure of the model's performance in each class.

1.4 Objective

The main goal of this thesis is to build a complete and standardized experiment for automated classification of scalp diseases using state-of-the-art deep learning architectures. This project is architecturally structured around seven chosen deep learning models. The selection of each model was based on the fact that it is a different stage or philosophy in modern computer vision architecture design. Images of scalp and hair disorders are frequently visually similar, have overlapping symptoms, and are imbalanced in terms of class distribution, making classification difficult. Hence, in this study, multiple architectures are tested under the same dataset, preprocessing, training and evaluation conditions to find out which model design is most effective for six-class scalp disease classification.

One of the other goals is to transform the Hair-Scalp-Analysis v2 dataset into a six-class classification problem. Cosmetic categories that are unrelated and redundant severity labels were eliminated to make the final data set more compact and consistent with the scope of our study.

The next goal is to train all seven models using a common experimental setup. The same image size, data split, optimizer, loss function, learning rate, scheduler, batch size and number of epochs are used to train each model. This will make sure the comparison is fair and the performance differences are primarily due to the architectural differences and not due to the different training conditions.

The ultimate goal is to perform a full quantitative assessment with several performance measures. This thesis uses Accuracy and macro-averaged Precision, Recall, Specificity, F1-score, and AUROC to evaluate each model as accuracy alone can be misleading in an imbalanced dataset. This enables the study to find not only the most accurate model, but also the one that has the best ratio of correct predictions of scalp disease classes to incorrect ones.

1.5 Thesis Organization

The remainder of the thesis is organized as follows:

- **Chapter 2 (Literature Review)** The existing work about scalp diseases, medical image classification, deep learning and disease detection using computer vision is reviewed. It reviews the difficulties in classification of scalp disease and explores the use of CNN based and transformer based architecture in medical image analysis. This chapter also outlines the research gap that this study aims to fill.
- **Chapter 3 (Dataset)** The Hair-Scalp-Analysis v2 dataset and the method for building the final six-class dataset are described here. It introduces the selected classes, distribution of the dataset, image preprocessing process, splitting strategy of the dataset and the data augmentation techniques used in the experiments.
- **Chapter 4 (Methodology)** The proposed experimental methodology is presented in this part. It outlines the seven deep learning architectures that were selected, the transfer learning approach, the custom classification heads, the training configuration, and the metrics for assessing the performance of the models.
- **Chapter 5 (Results and Analysis)** Experimental results of the baseline and augmented experiments are presented in this chapter. It offers quantitative comparison of the seven architectures using Accuracy, Precision, Recall, Specificity, F1-score and AUROC, class-wise performance analysis and relevant visualizations.
- **Chapter 6 (Conclusion & Future Work)** The research findings are summarized and conclusions are drawn in this chapter in light of the research questions. The chapter also considers the social impact, environmental impact and sustainability of the proposed approach and the ethical issues associated with it, before outlining potential directions for future research.

Chapter-2: Literature Review

Automated scalp diseases classification has received increasing research attention due to the advances in computer vision and deep learning. Various different approaches have been used for detecting and classifying scalp conditions using different neural network architecture, preprocessing techniques, and datasets.

2.1 Existing Research on Scalp and Hair Disorders

Many scalp and hair conditions are visually similar and difficult to identify especially by traditional visual inspection. This has inspired more research into machine learning and deep learning techniques that can learn discriminative visual features directly from scalp images. As a whole, the literature available shows a development from rather small-scale CNN-based classification systems and image-processing pipelines. One of the early works in this direction is by Roy and Protity [11] who studied the automated detection of three scalp and hair conditions namely alopecia, psoriasis and folliculitis. Their work focused on the significance of image preprocessing in the case of limited and heterogeneous image data, while later studies mainly compared the different established pretrained architectures. The authors gathered only 150 images from various sources on the Internet, and they denoised, enhanced, equalized and balanced the data before training a 2D CNN. The resulting model was able to achieve 96.2% training accuracy and 91.1% validation accuracy, while class-specific precision and recall values showed particularly good recognition of folliculitis. The main novelty of this work is thus not the comparison of complex architectures, but the fact that adequate preprocessing can significantly enhance the quality of the visual information that can be provided to a CNN when data are limited. However, the number of images in the dataset is small, the images are from various sources, and only three disease categories are included, which limits the generalizability of the results.

Chang et al. [9] took this idea a step further by creating an integrated system for scalp inspection and diagnosis using deep learning, named ScalpEye. Instead of using standard photos, ScalpEye used a portable scalp microscope with 200 times magnification, a mobile app, a cloud-based AI training server and a management platform. Four common scalp and hair symptoms were targeted in the system: dandruff, folliculitis, hair loss, and oily hair. Their approach was to view it as an object detection problem, not just an image-level classification problem. The three candidate architectures explored were: Faster R-CNN with Inception-v2, SSD with Inception-v2, and Faster R-CNN with Inception-ResNet-v2-Atrous. The final system achieved average precision ranging from 97.41% to 99.09% for the four symptoms using Faster R-CNN with Inception-ResNet-v2-Atrous. The system

could also estimate the severity of the symptoms and send the results via a cloud-based platform, which is an important strength of this study, as the system is moving towards practical deployment. This demonstrates that the utility of an AI diagnostic system is not just related to classification accuracy, but also image acquisition, usability, and integration into a practical workflow. But it is only applicable to a limited number of scalp-disease classes due to its design based on four symptoms and a special microscopic imaging device. The training data were just over 2,000 microscope images from a hair-care organization, so the model was trained under a rather specific imaging and acquisition environment. Furthermore, the system was developed for scalp-health and hair-care rather than for the full dermatological diagnosis, and thus significant disorders not included in the four target symptoms were not taken into account.

A more clinically oriented problem was tackled by Zhang et al. [4] who explored the ability to distinguish between two visually similar inflammatory scalp diseases, scalp psoriasis and seborrheic dermatitis. They analysed 1,358 dermoscopic images from 617 patients (508 with psoriasis and 850 with seborrheic dermatitis). The images were split into training, validation and testing sets, and the authors applied transfer learning using GoogLeNet, a 22-layer convolutional neural network that was pretrained on ImageNet. The goal here was not to classify multiple diseases, as in the overall multi-disease classification studies, but to differentiate between two diseases that might be hard to tell apart visually. The model was 96.1% sensitive, 88.2% specific, and had an AUC of 0.922 on the diagnostic task, and was said to perform better than the five dermatologists who were part of the comparison. Of note, the study also examined if AI support could enhance the performance of less-experienced clinicians, and the model boosted the AUCs of a dermatology graduate student and two general practitioners. This is especially relevant to the clinical motivation for automated scalp analysis as it shows that deep learning can not only be used as an autonomous classifier but also as a decision support tool for visually ambiguous conditions. However, its drawbacks are its restricted diagnostic problem and the fact that it is based on dermoscopic images obtained in a specific clinical setting. The model is only able to differentiate psoriasis from seborrheic dermatitis and can't be used to determine if the same model can be generalized to more scalp conditions that have similar visual features.

Kim et al. [10] then showed the value of model ensembles and data augmentation in the presence of a very limited scalps dataset. They specifically studied the classification of alopecia severity and tested the ResNet, ResNeXt, DenseNet and XceptionNet architectures along with ensemble combinations. The original alopecia images were very unevenly distributed with 13,156 images in the mild class, 3,742 images in the moderate class, 825 images in the severe class, and 526 images in the healthy class. To overcome these drawbacks, the authors used geometric augmentation, PCA-based colour augmentation, normalization and focal loss. DenseNet201 achieved the best average F1 score of all the individual networks, which was 77.12% without augmentation. ResNeXt101 had the highest average F1-score of 86.8% after augmentation. More significantly, the results were further enhanced by the ensemble configurations. The ensemble of ResNeXt101, DenseNet169 and XceptionNet41 model had the best F1 score of 87.74%, and DenseNet169 and XceptionNet41 model had the best accuracy of 95.84%. These results show that the interplay between dataset size, class imbalance, augmentation, and architecture may be more significant than just using the deepest model available. The study also shows that the fusion of architectures with complementary feature representations can enhance the performance and make class-wise predictions more balanced. The study is confined to alopecia, and the images of the scalp are microscopic, so that direct transfer to a larger set

of scalp images or to a set of ordinary photographs is not possible. The authors themselves identify the use of microscope imagery as a limitation and propose investigating images captured using regular cameras in future work. Therefore, the results of Kim et al. do not demonstrate which individual architecture is best suited to a wider classification problem of six or multiple classes for scalp diseases in a limited data scenario.

Ha et al. [2] extended the analysis of scalp conditions to six conditions: dryness, oiliness, erythema, folliculitis, dandruff, and hair loss, and added four severity levels for each condition. Their research pitted three very different deep-learning models against each other: ResNet-152, EfficientNet-B6, and ViT-B/16. The highest average classification accuracy was obtained by the ViT-B/16 model, which was 78.31%, compared to the two CNN-based models. The study is especially relevant because it is a shift from the traditional CNN-based classification to transformer-based image analysis, and because explainability was introduced in the form of attention rollout. Using the attention mechanism, the approximate lesion regions were identified without the need for additional lesion annotations, and the trained model was embedded in a software platform for scalp monitoring. The contribution goes beyond a classification score and takes into account interpretability and practical deployment, which are both relevant for medical AI. The relatively modest average accuracy of 78.31% however, reflects the challenge of severity-level classification, where the visual differences between adjacent severity levels are sometimes subtler than the differences between broad disease categories. The study relied on the AI-Hub scalp-image dataset which has a limited number of images, and lighting and visual variation of the scalp condition (oily, non-oily) can also impact prediction.

Nam Anh et al. [7] recently performed a direct comparative evaluation of five well-known CNN architectures: VGG16, VGG19, Inception-V3, ResNet50, and ResNet152 on a ten-class scalp and hair disease dataset. Their dataset comprised of about 12,000 images from Kaggle along with 530 images gathered from the Internet, and the classes were alopecia areata, contact dermatitis, folliculitis, head lice, lichen planus, male pattern baldness, psoriasis, seborrheic dermatitis, telogen effluvium, and tinea capitis. The data was split into training, validation and testing sets with 80:10:10 ratio. VGG16 achieved the highest test accuracy (96.81%), with VGG19 being the next (96.73%), and both models had precision, recall and F1-scores around 0.97. Inception-V3 also achieved a high accuracy of 95.13%, while ResNet50 and ResNet152 achieved only 49.16% and 36.87% respectively. These results are especially significant as they indicate that more architectural depth is not necessarily better in classification of scalp diseases. The authors state that the lack of diversity of the data set and the risk of overfitting may be the reason for the poor performance of the deeper ResNet models, in addition to the effect of hyperparameter selection. The study thus underscores the need to tailor models to the nature of the data set, and not necessarily to rely on newer or deeper models. The main drawback is that, despite the seemingly large number of images, the diversity of the dataset is rather small. The authors point out that the dataset only includes 10 common conditions and doesn't include many of the more rare diseases. They consequently recommend expanding the dataset with more diverse images and conditions and further investigating ensemble and attention-based methods. The study is of particular value for the present research because it offers direct evidence that comparative benchmarking among several architectures is required, given that no single architecture can be assumed to be the best for the processing of scalp imagery.

The link between architecture selection, domain specific training and diagnostic perfor-

mance is further supported by evidence from other fields of medical imaging. Bhuiyan et al. [13], for example, compared six deep-learning models ranging from CheXNet, DenseNet201, DenseNet121, ResNet101, EfficientNetB3, to MobileNetV2 for a five-class chest X-ray classification involving normal cases, COVID-19, viral pneumonia, tuberculosis, and lung opacity. CheXNet had the best performance with an AUROC of 0.988 and accuracy of 94.12%, followed closely by EfficientNetB3 with an AUROC of 0.987 and accuracy of 93.35%. The lowest AUROC of 0.963 was obtained by MobileNetV2. A specific finding of particular relevance is that CheXNet's benefit was linked to its medical-image pretraining, which means that the utility of an architecture is not just a function of its architecture, but also of the knowledge that is captured in its pretrained representations. Grad-CAM was also adopted to interpret the model predictions by visualizing them. While this study is on chest radiography and not on scalp disease, it is relevant because it indicates that the same general CNN principles can be successfully transferred from one medical imaging domain to another and that the choice of domain-specific pretraining and architecture can have a tangible impact on performance. However, its results should not be directly compared numerically to scalp studies, since the image modality, anatomical structure, disease presentation and scale of the datasets are very different between chest X-rays and scalp photographs.

Yet there is a significant lack in the literature. Most studies that are specific to the scalp are limited to either a few diseases, or one disease, or estimating severity, or a special imaging modality. Roy and Protity [11] only considered three conditions; ScalpEye [9] considered four symptoms using a dedicated microscopic device; Zhang et al. [4] focused on two diseases, psoriasis and seborrheic dermatitis, using dermoscopy; Kim et al. [10] focused specifically on alopecia severity; and Ha et al. [2] considered six conditions but emphasized severity-level classification rather than broad disease discrimination. The comparative study by Nam Anh et al. [7] is more extensive with 10 classes, but the dataset is heterogeneous, including head lice and cosmetic/hair-related diseases in addition to dermatological scalp diseases, and the study mainly compares a set of existing CNN architectures.

2.2 Challenges in Scalp Disease Classification

The classification of scalp disease is a difficult task due to the presence of numerous disorders with similar appearance to each other [3][4]. Scaling, redness, follicular inflammation, and thinning of hair may manifest in many conditions [2][3]. This overlap renders visual classification challenging even in trained observers and it promotes misclassification in automated systems too. Scalp images also have fine-grained local cues that the model should learn to identify but the irrelevant variation that should be ignored includes hair color, hairstyle, image background and cosmetic difference. This need is biased towards architectures able to simultaneously capture local texture features as well as larger contextual features in a single image [18].

Dataset reliability is also a concern. Datasets that are publicly available can be used in benchmarking, but they can be noisy labeled, have uneven lighting, differing camera angles, and may be taken outside of the clinical environment of the study [10]. These aspects may decrease the quality of the trained model when it is used in real-world

Figure 3: An image showing that various scalp diseases have visual overlapping features [7].

situations [3]. A well-performing model on one publicly available dataset might not be well-performing on images taken with other cameras, in varied lighting environments, with diverse demographic groups, or with alternative clinical tools [3]. False positives and false negatives should also be carefully considered in the clinical use. False positive can result in the unnecessary worry or further examination not required whereas false negative can result in postponed diagnosis and reduced treatment. This is why recall and specificity should be discussed in conjunction, particularly in cases where the system is supposed to help in clinical screening. Another practical issue is the efficiency of the computations. A model that is to be used on a mobile device or in a screening situation in real time might need to be operated on a low processing power hardware. This requirement can be met by lightweight architectures, although they need to provide reasonable accuracy and predictable performance as well [9][14]. This is among the reasons why this study has a comparison of heavy, efficient and transformer based model types.

Another difficulty with scalp disease classification is that of distinguishing between the various types of hair loss that can have similar visual characteristics. Interestingly, this study has distinguished between the two conditions, Alopecia Areata (Alopecia) and Androgenic Alopecia (Hair Loss). It can be difficult to distinguish the two from this study, particularly when both involve exposed scalp or reductions in hair density. Diseases have different clinical patterns and underlying causes, but variations in the process of disease, severity, and individual appearance can lead to images with similar visual characteristics. The ability to differentiate between the various forms of alopecia has also been noted as difficult in previous studies, especially when certain clinical features are similar, or when images were insufficient for diagnosis. This is a significant challenge for automated image-classification systems, because similar conditions can result in misclassification of the image, even though they are actually different types of clinical conditions. The problem is of particular interest to this study because the two classes (Alopecia vs. Hair Loss) are conceptually related and share some common visual characteristics, thereby reflecting the fact that the definition and makeup of disease classes can directly affect classification accuracy.

2.3 Scope of Study

This thesis only covers the image-level classification of 6 scalp and hair conditions using deep learning. The classification of each model is the class label predicted from the following 6 classes: Healthy Scalp, Alopecia, Folliculitis, Dermatitis, Dandruff, Hair Loss. A single refined dataset is used for all models, which is taken from the Hair Scalp Analysis v2 repository on Roboflow Universe [19]. There is no external dataset used for validation. Therefore, the results should be interpreted as an internal validation and not as a complete clinical validation. Comparative architectural performance is the core of the study. Transfer learning is used and the same evaluation metrics are applied to all models, which are trained under the same conditions. This uniform configuration enables the study to determine which architectures are most suitable for the task and the constraints of the dataset.

Chapter-3: Dataset

The quality, diversity and organization of data used for training and evaluation are very critical to the performance of any deep learning model. Given that this study aims to classify scalp diseases automatically from image data, the selection of the dataset and preprocessing were crucial to ensure that the model learns meaningful disease-specific features while reducing the impact of irrelevant or redundant classes. The main dataset utilized in this research is described in this chapter, as well as the modifications made to suit the study objectives, the refining procedures applied and the augmentations applied on a separate set prior to model training. The results of these experiments were compared.

3.1 Dataset: Hair-Scalp-Analysis v2

For this study, the dataset was sourced from the publicly available Hair-Scalp-Analysis v2 repository on Roboflow Universe [19]. The original repository had 3544 JPG images of three channels (RGB) and 640 x 640 resolution. These images were spread out over sixteen classes: Alopecia, Dandruff, Dermatitis, Dry Hair, Folliculitis, Grey Hair, Hair Loss, Hair Loss Stage 1, Hair Loss Stage 2, Hair Loss Stage 3, Hair Loss Stage 4, Head Lice, Healthy Scalp, Oily Hair, Psoriasis, and Red Hair.

3.2 Class Selection and Dataset Refinement

Both disease and non-disease classes were included in the original classes. Some classes were cosmetic conditions like Grey Hair, Red Hair, Oily Hair. Another class, Head Lice, was also eliminated, because it is not a disease of the scalp itself, but rather a parasitic infection caused by Pediculus humanus capitis, and there were few samples in the original dataset. There were also several Hair Loss Stages, which added redundancy since there was already a Hair Loss class. Unrelated cosmetic classes and redundant severity-stage labels were deleted to make the dataset more relevant to the planned research objective. The final dataset comprised of six classes of target namely Healthy Scalp, Alopecia, Folliculitis, Dermatitis, Dandruff, Hair Loss. To highlight, the Alopecia class is also known as Alopecia Areata, which is characterized by patches of hair loss and hair thinning across the entire scalp. On the other hand, Hair Loss is associated with hereditary and genetic factors of hair thinning, baldness and a receding hairline, also known as Androgenic Alopecia. This refinement resulted in a focused disease classification problem with a healthy control class.

There were 1368 images in this dataset. Healthy Scalp was the most common class and Hair Loss was the least common class. This is significant for interpretation as a model can be highly accurate on the majority classes and under-performing on the minority classes. Therefore, the evaluation framework incorporates measures other than accuracy. This class refinement is summarized in the form of splits in Table 3.1. Some samples of the images from the rest of the classes after cleaning the dataset are shown in Figure 4.

Table 3.1: Summary of class distribution in the final dataset.

Class	Number of Images	Train Split	Validation Split	Test Split
Healthy Scalp	515	360	78	77
Alopecia	245	171	37	37
Folliculitis	212	148	32	32
Dermatitis	200	140	30	30
Dandruff	114	80	17	17
Hair Loss	82	56	13	13
Total	**1368**	**955**	**207**	**206**

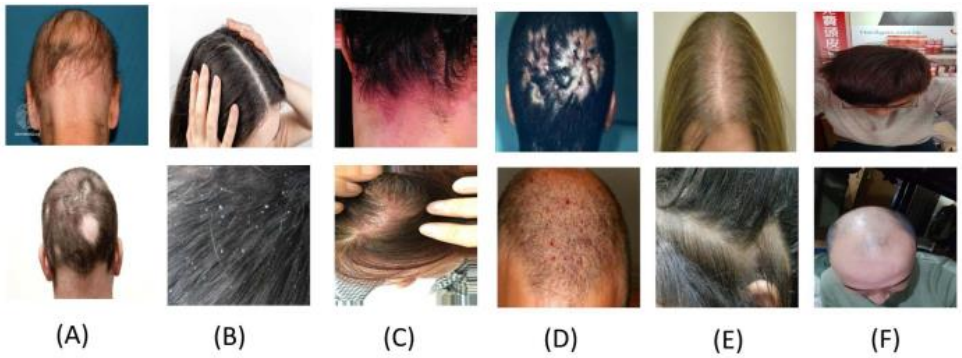

Figure 4: Dataset display with classes (A) Alopecia, (B) Dandruff, (C) Dermatitis, (D) Folliculitis, (E) Hair Loss, and (F) Healthy Scalp.

3.3 Data Splitting

The filtered dataset was further split into training, validation, and testing sets in the ratio of 70%, 15%, and 15%. This approach enables the model to be trained with the training set, optimised for performance with the validation set, and then evaluated with the testing set. Fair comparison is provided by a consistent split across models. All images were resized to 224 x 224 pixels before being fed to the network for training. This dimension was chosen, as the backbones used for evaluation were pre-trained on ImageNet and usually take in images of a similar input size. Resizing also helps to align the shape of the input tensor, so that multiple architectures can be trained along the same pipeline.

3.4 Data Augmentations

The geometric augmentations in this study involve random rotation, horizontal and vertical flipping, translation, scaling, shearing, affine transformation, perspective transformation and cropping the sample of the dataset. The samples are also modified in color and intensity through various methods including brightness changes, contrast changes, saturation changes, hue changes, color jitterring or gamma corrections. An augmented version of the scalp disease dataset was created to enhance the generalisation ability of the deep learning models and to minimise the impact of class imbalance. Following multi-label filtering, the final dataset consisted of 1368 images that are classified into six classes: Alopecia, Dandruff, Dermatitis, Folliculitis, Hair Loss, and Healthy Scalp. The image distribution of classes was skewed with Healthy Scalp having the highest number of images at 515 and Hair Loss the lowest number of images at 82.

To enhance the diversity of the images and maintain their medical relevance, several augmentation techniques were used. Random rotation, random zoom or crop, vertical flip, horizontal flip, affine transformation, brightness adjustment, and contrast adjustment. The images were randomly rotated between 90° to 90°. Random zoom or crop was applied discretely (not a continuous value between a range) with a magnitude of 20%. By default random vertical and horizontal flips are discrete. The brightness and contrast adjustments were applied with random affine transformations of magnitude from 0% to 10%, however; each of the random affine transformations was applied to the brightness and contrast with a continuous magnitude between 0% and 15%. As can be seen in Table 3.2 the type of augmentation, along with its probability and the magnitude of its application is shown.

Table 3.2: Augmentation techniques used on the dataset.

Augmentation Type	**Probability**	**Magnitude**
Random Rotation	0.3	90° to −90°
Random Zoom/Crop	0.3	20%
Random Vertical Flip	0.3	–
Random Horizontal Flip	0.3	–
Random Affine Transform	0.2	10%
Random Brightness	0.2	15%
Random Contrast	0.2	15%

The number of samples for each class was used to determine the augmentation factor. Lower numbers of samples were supplemented more. In the Hair Loss class, for instance, there were originally only 82 images after filtering, and the authors augmented this class of images by five, so a training set of 56 images was increased to 336 images. Likewise, the number of images in the Dandruff class was increased from 80 to 320 by using an augmentation factor of 3. Alopecia, Dermatitis and Folliculitis were moderately increased, and the Healthy Scalp class was not increased due to the highest amount of samples already. The new number of images for the train, validation, and test sets, respectively, was 2,347.

Only the training subset was enhanced with data augmentation. Validation and test subsets remained the same to avoid bias in the evaluation. The models were thus trained with higher variations in the appearance, visual angle and cropping of the scalp images,

as well as in the lighting of the image. This enhanced the robustness of the classification process and helped to minimize overfitting, particularly for minority classes like Hair Loss and Dandruff. Table 3.3 shows the distribution of the data after it is augmented, and Figure 5 shows some sample examples of the images after they are augmented with a single augmentation technique and multiple augmentation techniques.

Table 3.3: Summary of class distribution in the augmented dataset.

Class	**Number of Images**	**Train Split**	**Validation Split**	**Test Split**
Healthy Scalp	515	360	78	77
Alopecia	416	342	37	37
Folliculitis	360	296	32	32
Dermatitis	340	280	30	30
Dandruff	354	320	17	17
Hair Loss	362	336	13	13
Total	**2347**	**1934**	**207**	**206**

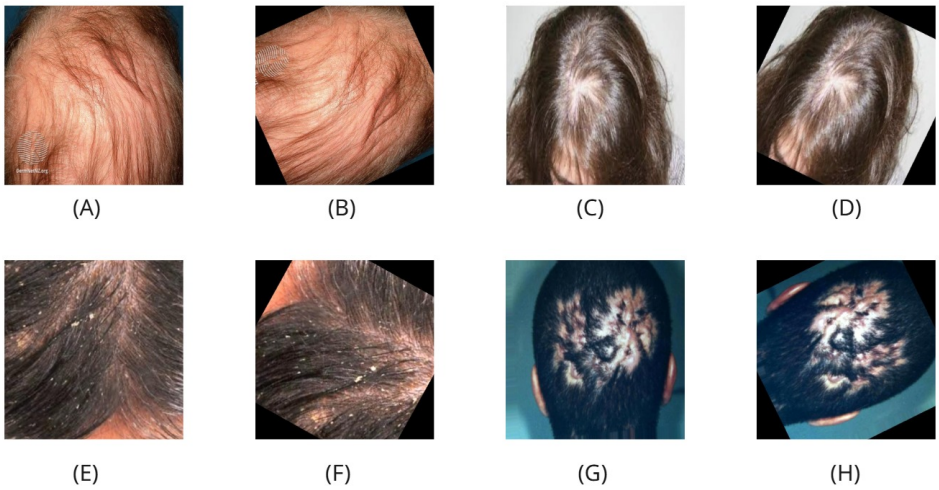

Figure 5: Augmented Dataset display with instances (A) Raw Image instance 1, (B) Translated + Rotated Image, (C) Raw Image instance 2, (D) Translated + Rotated Image 2, (E) Raw Image instance 3, (F) Rotated + Contrasted Image, (G) Raw Image instance 4, (H) Rotated + Brighter Image.

Chapter-4: Methodology

The proposed method follows a supervised image classification pipeline. A pretrained model backbone is loaded and modified with a classifier head suitable for six-class prediction. Finally, the trained model is evaluated using multiple performance metrics.

As shown in Figure 6, the process begins with the acquisition of scalp images from the prepared dataset. The dataset is resized and then divided into training, validation, and testing subsets to facilitate model learning, hyperparameter tuning, and unbiased performance evaluation. The processed images are then supplied to a selected feature extraction backbone. Rather than training a network from scratch, this study employs transfer learning by utilizing pre-trained CNNs and a Vision Transformer (ViT) model. These backbone networks automatically learn hierarchical visual features from the input scalp images, including low-level characteristics such as edges and textures, as well as high-level disease-specific patterns.

The extracted feature representations are subsequently passed to a custom classifier head, illustrated in the green block of Figure 6(A). The classifier consists of two fully connected (dense) layers containing 1024 and 512 neurons, respectively, with each layer followed by a Rectified Linear Unit (ReLU) activation function. These layers progressively transform the extracted feature vectors into a more discriminative representation suitable for classification. A Flatten layer is employed to convert multidimensional feature maps into a one-dimensional feature vector before the final classification stage. Finally, the output layer contains six neurons, corresponding to the six scalp condition categories, and produces the predicted class probabilities.

4.1 Model Overview

An effort was made to identify seven architectures to illustrate various deep learning design philosophies: DenseNet121, ResNet34, ResNet50, EfficientNetB0, EfficientNetB2, ConvNeXt-B, and ViT-B/16. Six of these models is CNN-based and ViT-B/16 is a transformer-based model. This choice allows a comparative analysis among densely connected CNNs, residual networks, efficient CNNs, modernized CNN architectures, and transformer-based vision models, which are separated by different stages and philosophies in designing modern computer vision architectures as illustrated in Figure 6(B). The models were all initialized with ImageNet-1K pretrained weights, which enabled them to utilize the visual feature representations that have been previously learned. To fit the architectures to the need of this study the initial classification layers have been substituted with new

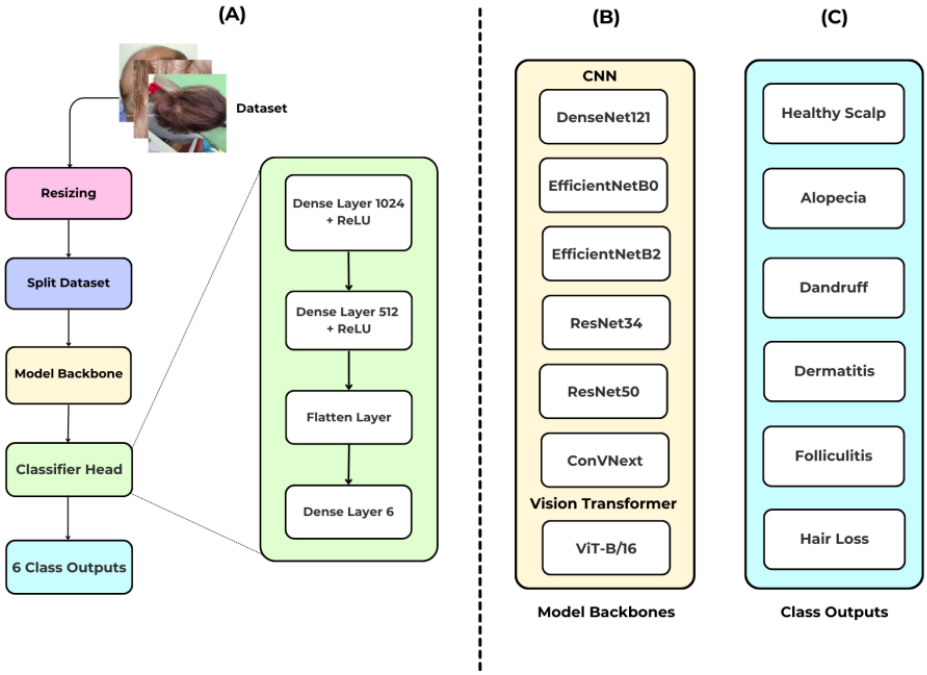

Figure 6: Methodology flowchart showing (A) system pipeline, (B) model backbones, and (C) class outputs.

classification heads that are aimed to generate six output classes of the desired scalp conditions. The backbone parameters were frozen with the new classification heads being trained on the scalp disease data. This transfer learning approach lowers computation and data cost to learn low-level and general visual features by direct training, and is therefore especially appropriate to the fairly small size of the medical image data set utilized in this experiment.

4.1.1 DenseNet121

DenseNet121 is a deep convolutional neural network (CNN) architecture that was introduced by Huang et al. to solve various shortcomings of traditional CNNs, such as vanishing gradients, redundancy of features, and poor use of parameters. DenseNet, unlike traditional convolutional networks, does not rely on the immediately previous layer to provide inputs to a layer, but instead provides a direct connection between every layer with all the following layers of the same dense block. The input to any layer is therefore the combined feature maps of all the previous layers, this allows the very large reuse of features and enhanced information flow permeates the entire network. The name of DenseNet121 is formed based on the number of learnable layers (121) and adheres to a hierarchical structure including an initial convolutional layer, four dense blocks, three transition layers, and a final classification layer. The major benefit of DenseNet121 is that it helps to resolve the problem of vanishing gradient. The high-performance of denseNet121 on a wide range of medical image classification tasks has resulted from its capability to retain fine-grained visual details with a relatively low level of computational complexity. Past works have effectively used dense net121 to classify skin diseases, retinal cases, chest radiographies and histopathological images where minute differences in texture are important to achieve correct diagnosis. The dense feature propagation methodology is specifically useful in the analysis of dermatological images since ailments of the scalp can have a great deal of

visual similarity that cannot be discriminated by low level texture clues alone but instead necessitates high level semantic variable clues. Figure 7 displays a simplified architecture diagram of the DenseNet121.

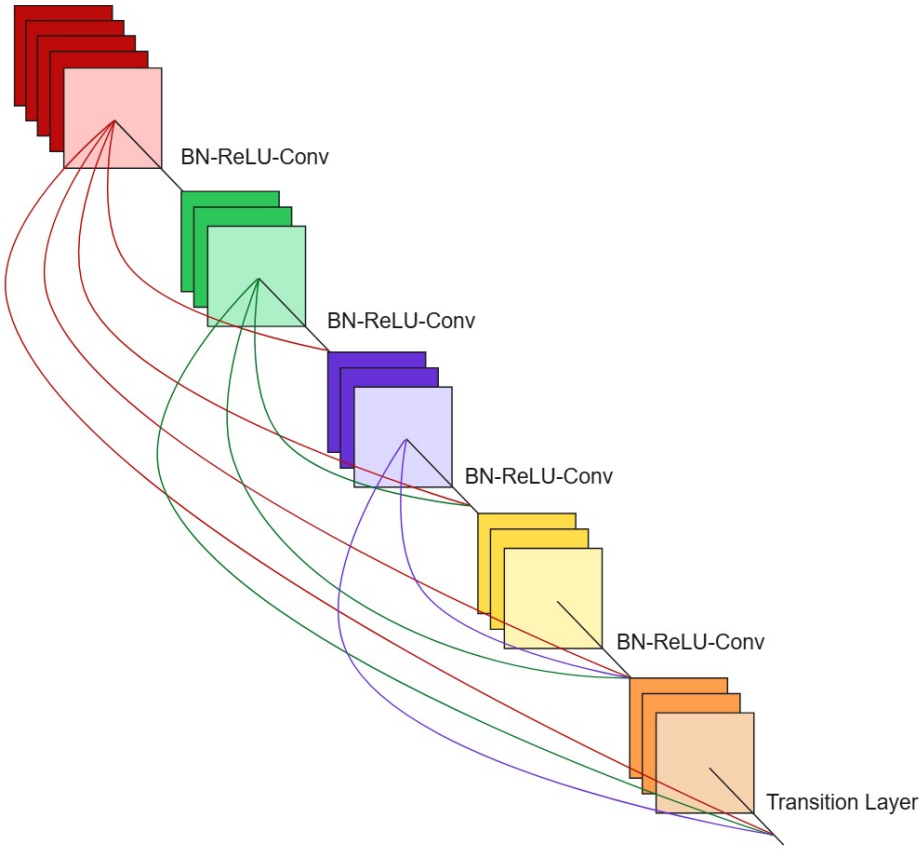

Figure 7: Simplified architecture diagram of the DenseNet121 [17].

4.1.2 ResNet34 and ResNet50

Residual Network (ResNet) is a deep convolutional neural network design that was proposed by He et al. to address the degradation issue that occurs when training very deep neural networks. Until the advent of ResNet, adding more depth to a network tended to enhance its feature representation, but past a certain threshold deeper networks tended to have higher training error and worse accuracy than shallower ones. Overfitting did not cause this degradation, but rather optimization challenges that did not allow the network to learn the desired mapping effectively. ResNet suggested the concept of residual learning to overcome this obstacle so that deep networks can be trained effectively and still perform well. An average residual block comprises of several convolution layers with Batch Normalization (BN) and Rectified Linear Unit (ReLU) activation functions on each side. The shortcut connection is added after the convolutional operations to the processed output and then a final activation function is applied to the sum of the original input and the processed output. When the input and output dimension of features are not equal, the shortcut path uses a 1–1 convolutional projection in order to make the dimensions compatible before adding the elements. This architecture maintains valuable low-level image information and at the same time learns more and more complex features representations. ResNet has proven to be effective in computer vision and the analysis of medical images. The initial ResNet design

set state-of-the-art record on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by being able to train networks up to 152 layers with a lower computational complexity compared to prior designs like VGGNet. Follow-up studies also indicated that identity shortcut links contribute to faster information dissemination and optimization in the extremely deep residual net networks [20]. These benefits have seen ResNet widely be utilized in a number of medical imaging tasks, such as the classification of skin lesions, retinal disease, the analysis of chest X-rays, and the classification of histopathological images, in which the quality of feature extraction is vital to differentiate between visually similar patterns of disease. Figure 8 shows two versions of the ResNet34 architecture.

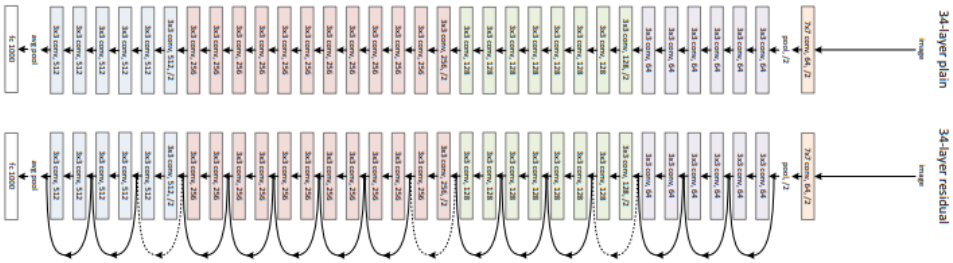

Figure 8: Architecture diagram of the ResNet34, plain and residual versions [18].

4.1.3 EfficientNetB0 and EfficientNetB2

EfficientNet A set of convolutional neural net (CNN) architectures suggested by Tan and Le to enhance classification accuracy without too much computational cost. Contrary to traditional CNN structures, where network depth, network width, or input image resolution are independently scaled, EfficientNet presents a compound scaling technique, which uniformly scales all three dimensions by a predefined set of scaling coefficients. With this balanced scaling approach, the network can attain a high level of performance using a much smaller number of parameters and lower computational cost than the state-of-the-art architectures of the time, like ResNet and DenseNet. The EfficientNet architecture starts with the baseline model, EfficientNet-B0, that was automatically learned with Neural Architecture Search (NAS). Instead of manually making the network, NAS will be used to optimize the architecture to achieve the highest classification accuracy at the lowest cost of computation. The resulting EfficientNet-B0 architecture is composed mainly of Mobile Inverted Bottleneck Convolution (MBConv) blocks which were first presented in MobileNetV2 alongside squeeze-and-excitation (SE) attention modules which dynamically re-calibrate channel-wise feature responses. These low-weight building blocks allow EfficientNet to learn very discriminative image features at computational efficiency [21]. Figure 9 demonstrates an architecture diagram of an EfficientNet with the scaling process that is based on the baseline.

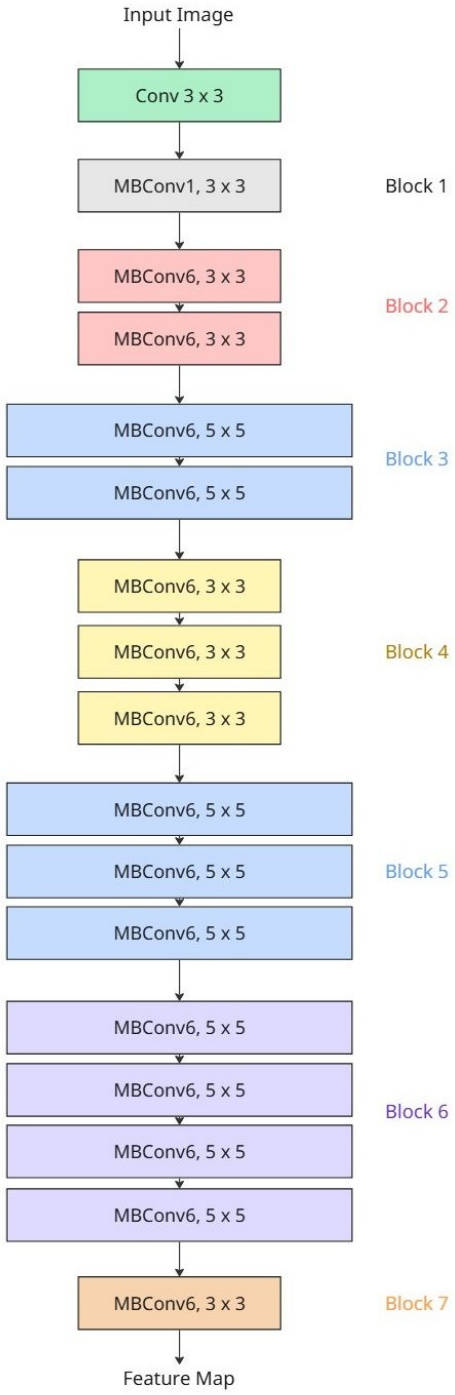

Figure 9: Architecture diagram of an EfficientNet with the baseline scaling process [14].

4.1.4 ConvNeXt-B

ConvNeXt-B ConvNeXt-B is a contemporary convolutional neural network (CNN) architecture suggested by Liu et al. that reevaluates and reimagines conventional convolutional networks using a number of architectural design principles based on Vision Transformers (ViTs). Although Vision Transformers showed impressive results when trained on large scale image recognition tasks, they are frequently much larger in terms of data size and

computation. ConvNeXt-B was created to explore whether traditional CNNs could obtain similar results through a systematic exploration of modern design methods coupled with maintaining the efficiency and inductive biases of convolutional operations. ConvNeXt-B was developed based on the ResNet architecture, which was the initial network. Instead of proposing a completely new architecture, Liu et al. continuously transformed ResNet with a number of innovations based on the recent developments in transformer-based vision models. These changes involved the substitution of the classic bottlenecked residual blocks with inverted bottleneck designs, scaling up the convolution kernel sizes to 7x7, minimizing the quantities of activation and normalization layers, substituting Batch Normalization (BN) with Layer Normalization (LN), and adding individual downsampling layers between network layers. With these progressive improvements, ConvNeXt-B was able to achieve substantial improvements in classification with minimal complexity and computational cost of convolutional networks. ConvNeXt-B was found to have competitive accuracy on various computer vision tasks, such as image classification, object detection, and semantic segmentation. ConvNeXt-B has been shown to be quite effective in medical image analysis because it can capture fine-grained local textures as well as larger anatomical details. Recent papers have managed to apply ConvNeXt-B to tasks like skin lesion classification, diagnosis of retinal disease, interpretation of chest X-rays, and analysis of histopathological images with reported better performance than conventional convolutional architectures in several studies [22] [21]. ConvNeXt-B is especially well suited to dermatological image classification due to the large receptive fields, the ability to extract features efficiently, and the strong optimization properties of large receptive fields, where subtle changes in scalp texture and lesion edges and inflammatory patterns are vital to differentiate between disease categories. The baseline ConvNeXt-B is shown as the diagram in Figure 10.

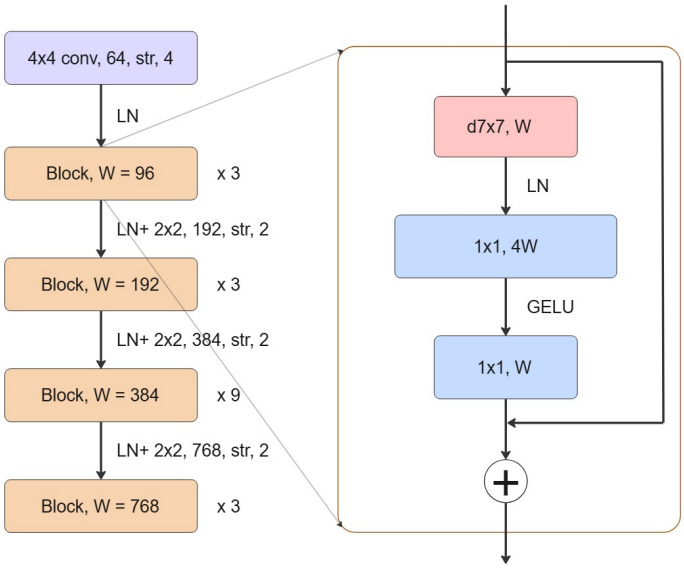

Figure 10: Simplified architecture diagram of the ConvNeXt-B with a block [15].

4.1.5 ViT-B/16

A transformer-based image classification model named the Vision Transformer (ViT) proposed by Dosovitskiy et al. is a Transformer architecture that is directly applied to image recognition problems. In comparison to conventional CNNs, where visual features

are computed by convolutional operations, Vision Transformers encode an image as a series of fixed-size patches that are computed in the same way as word tokens in a sentence. The new methodology allows the model to learn long-range correlations between various parts of the scalp image with self attention instead of local convolutional filters. The basic idea of a Vision Transformer is to replace an input image with a series of non-overlapping image patches. The ViT-B/16 architecture that will be used in this research separates each input image into 16 × 16 pixel patches with the notation B/16 representing the Base model with a patch size of 16 pixels. For an input image of size 224×224, this results in 196 image patches (14×14). Every patch is reduced to a one-dimensional vector and transformed to a fixed-dimensional embedding via a trainable linear projection layer. Just like token embeddings in NLP, positional embeddings are learned together with each patch embedding to maintain spatial information, which is otherwise lost in the flattening operation. Patch embedding is followed by the addition of a special classification token (CLS token) to the sequence and then it is fed into the Transformer encoder. The encoder is multiple identical layers each of which is composed of Multi-Head Self-Attention (MHSA), Layer Normalization (LN), Multi-Layer Perceptrons (MLPs) and residual skip connections. In the self-attention mechanism, every patch in an image attends to all the other patches in the image, such that the network can model the global relationships between contextual relationships in the image irrespective of the distance between them. This is in contrast to CNNs where receptive fields are expanded in a slow way by means of repeated convolutional layers. The final encoder layer output that represents the CLS token can be used as the global image representation and then be classified. Figure 11 shows a process diagram of the Vision Transformer.

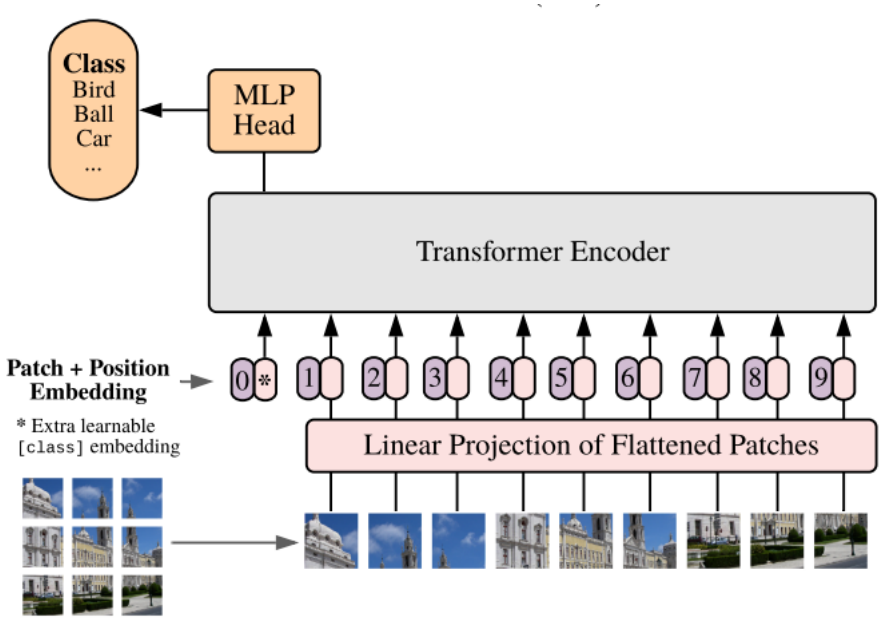

Figure 11: Architecture diagram of the Vision Transformer [16].

4.2 Training Configuration

All models were trained using PyTorch with the same training configuration. The models used CrossEntropyLoss as the loss function and Adam as the optimizer with a learning rate of 0.0001. A ReduceLROnPlateau scheduler monitored validation loss and reduced

the learning rate when improvement slowed. Each model was trained for 50 epochs with a batch size of 32. The training was performed using Google Colab with a Tesla T4 GPU. The same computational environment and hyperparameter choices were maintained across models to ensure fairness. Table 4.1 shows a summary of the model's training configurations.

Table 4.1: Summary of model training configuration.

Parameter	**Value**
Framework	PyTorch
Pretraining	ImageNet-1K
Loss Function	CrossEntropyLoss
Optimizer	Adam
Learning Rate	0.0001
Scheduler	ReduceLROnPlateau
Epochs	50
Batch Size	32
Hardware	Google Colab Tesla T4 GPU

4.3 Evaluation Metrics

We selected four basic components which we evaluated as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

- **True Positive (TP)** is the number of images of a specific class of scalp diseases that are correctly classified as that class. For instance, if an image of Alopecia is correctly classified as Alopecia, it adds one to the True Positive value of the Alopecia class. A higher TP value means that the model is good at detecting the target disease.
- **True Negative (TN)** The number of images that do not belong to the target class and are correctly predicted as other classes is called True Negative (TN). For example, all images that are part of the Alopecia class and are correctly classified as not being Alopecia count towards the True Negative count. A high TN value means that the model has a good ability to distinguish the target disease from all other scalp diseases.
- **False Positive (FP)** are the images which are not of the target class but are predicted as belonging to the target class. For instance, if a Dandruff image is incorrectly classified as Alopecia, it counts towards the False Positive of the Alopecia class. False Positives are false alarms, when the model says there is a disease when there is not.
- **False Negative (FN)** are images that are actually of the target class but are misclassified as other disease class. For instance, if the image of Alopecia is predicted as Hair Loss, it is counted as False Negative for Alopecia class. False Negatives are instances where the model predicts a negative outcome when the truth is positive, which is especially critical in medical diagnosis as undiagnosed diseases can result in delayed treatment.

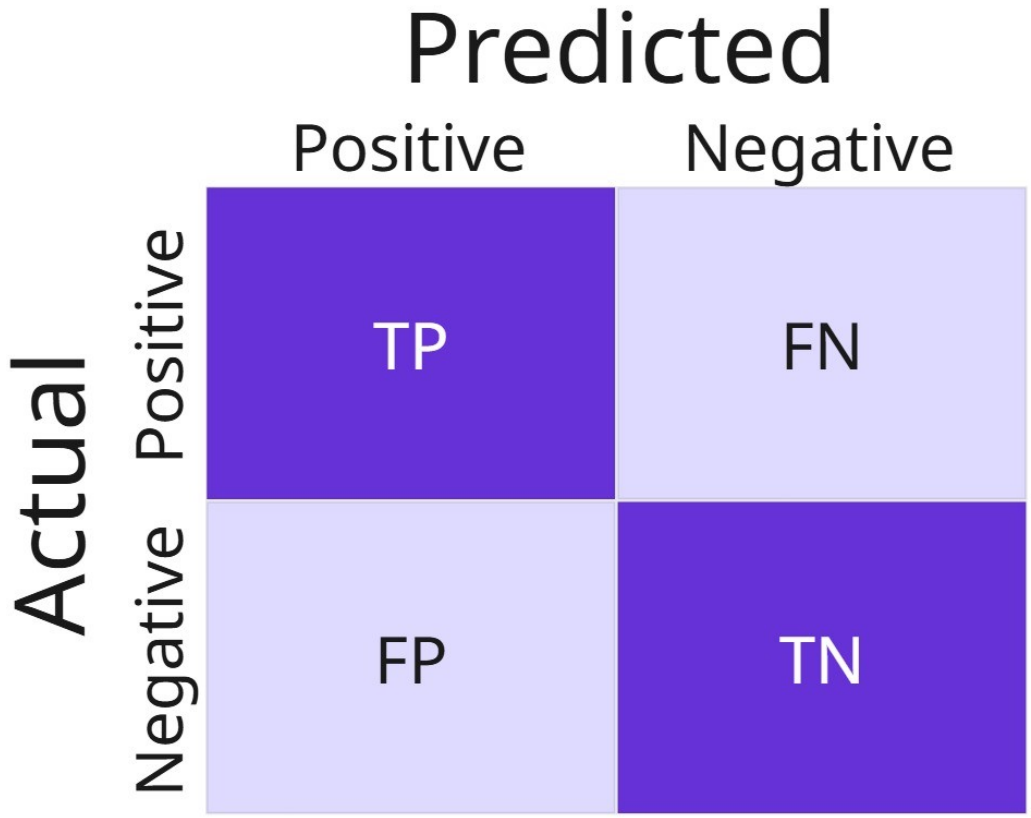

Figure 12: Diagram of the base confusion matrix, illustrating the layout of the Positive and Negative cases along with the Actual and Predicted cases.

There are six disease categories, but the evaluation metrics are evaluated one at a time – one-versus-rest. One disease class is used as the positive class and all other classes are combined into one negative class. This allows computing for each class its own values for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values, and afterward calculating the mean values across all of these in order to get a single average per class value. The four quantities described in the following sections are directly related to each of the evaluation metrics described below.

4.3.1 Accuracy

The accuracy represents the percentage of images that are correctly classified, across all the test images. It constitutes the percentage of the six classes' images that were correctly identified as being that class in this study.

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

Considering the need for accuracy is helpful, since it provides a direct indication of the total classification performance of the model. But sole accuracy can be deceiving in the case of an imbalanced data set. For instance, one class may contain much more images than another, and so a model that achieves high accuracy may do so on the more dominant

classes and a poor performance on the majority classes. Hence, in this thesis, the accuracy measurement is provided along with other measurements to ensure a better measurement of performance.

4.3.2 Precision

Precision is defined as the proportion of samples that were correctly classified as a certain class. It is about how the model's good predictions are likely to turn out. From the point of view of the classification problem of scalp disease, high precision means that if you give the model a classification of "alopecia" or "dermatitis," then it is likely you will get it right.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Precision becomes crucial when a false positive number must be kept to a minimum. For instance, if a healthy scalp picture is mistaken for diseased, it could result in anxiety or additional evaluations. Due to the use of multiple classes this study has made precision calculations for each class which are then subsequently macro-averaged to ensure equal weighting across disease categories.

4.3.3 Recall

Recall is the number of true positive observations correctly predicted by the model. It is also referred to as "sensitivity" or simply "sense". The present study is based on the premise that recall represents the ability of the model to identify true incidents of each type of scalp disease.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

A high recall is crucial for medical image classification due to the increased risk of the opportunity cost of a false alarm, versus a missed true case. For instance, an image is in the Hair Loss class but the model tags it as Healthy Scalp or Dandruff, in which case, the model fails to recognize this class. Thus, a key metric of a model's performance in correctly identifying patients with disease is recall.

4.3.4 Specificity

Specificity is how well the model performs in correctly identifying negative cases. That is, it measures the model's performance at rejecting legitimate samples that are outside of the class under consideration.

$$Specificity = \frac{TN}{TN + FP} \tag{4.4}$$

The higher the specificity the fewer false positive predictions the model will make. Specificity allows for less false-positive diagnosis of individual disease classes when defining the classification of scalp diseases. However, there may be a model with low specificity for Dandruff which would consequently classify many non-Dandruff images as Dandruff. Thus, specificity gives insight into the vigilance and/or discriminatory capabilities of the model to be used for misclassified class labels.

4.3.5 F1-Score

F1 score is the harmonic mean of the precision and the recall. It measures the balance between accuracy (correct positive forecasts) and the ability of the test to identify the true positive test results.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.5}$$

The F1-score is especially convenient when the data set is imbalanced. A model can be very accurate but have low recall, that is, it can correctly predict a class, but miss a lot of true samples of that class. But a large number of false positives are acceptable if the recall is high while precision is low, that is, if the model has a false alarm rate that is high. The F1-score is a compromise between these two.

However, in this thesis the F1 score is one of the important values, since there are a lower number of samples in some classes of the image (e.g. Hair Loss and Dandruff) compared to the rest of the larger classes (e.g. Healthy Scalp). Thus the F1-score is able to illuminate if the model works consistently for all classes or just works very well for most of the classes.

4.3.6 AUROC

The term AUROC is the acronym for Area Under the Receiver Operating Characteristic Curve. Determines the performance level of a model to classify a class from other classes on various decision thresholds. The ROC curve is plotted with the True Positive Rate (TPR) on the Y-axis and the False Positive Rate (FPR) on the X-axis and is used to measure the separability performance of the model. The larger the AUROC value the better the model is able to separate the classes. Values that are closer to 1.0 will show better ability to separate the classes while those close to 0.5 are less able to differentiate or there is a random classification ability.

To assess the overall class separability for each model the AUROC has been used in this work. This is to be noted, as two models could be equally accurate, but still differ in their ability to class the different classes of diseases. Because AUROC is a threshold-independent

measure, it is an appropriate metric for gauging a model's performance relative to its level of confidence for differentiating between various categories of scalp disease.

$$AUROC = \sum_{i=1}^{n-1} \frac{TPR_i + TPR_{i+1}}{2} (FPR_{i+1} - FPR_i) \tag{4.6}$$

$$TPR = \frac{TP}{TP + FN} \tag{4.7}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.8}$$

TPR_i is the True Positive Ratio at the i^{th} ROC operating point, i.e., its ability to recognize a true positive when there is one. The True Positive Rate is the next ROC operating point in the detections count; TPR_{i+1}. FPR_i is the False Positive Rate at i^{th} ROC operating point, percentage of actual negative that get misclassified as positive. The False Positive Rate of the next ROC operating point is called FPR_{i+1}. i refers to the current index of the ROC operating point and n is the number of all ROC operating points utilized to get the ROC curve. The rate of True Positive (TP) vs sum of False Positive (FP) and True Negative (TN) cases provides the True Positive Rate and the rate of False Positive (FP) vs sum of True Negative (TN) and False Positive (FP) cases provides the False Positive Rate.

Chapter-5: Results and Analysis

The seven models were trained on the original (baseline) dataset then trained with data augmentations (augmented). After each training session, all models were evaluated using the selected metrics. Their results are elaborated below.

5.1 Baseline Results

The baseline evaluation results show that ConvNeXt-B and EfficientNetB2 achieved the highest accuracy of 92.23%. However, ConvNeXt-B achieved a stronger balance across precision, recall, and F1-score, making it the most effective overall model in this study. Table 5.1 is a tabulation of the results.

Table 5.1: Evaluation results of the models.

Model Name	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**	**AUROC**
DenseNet121	89.81%	80.30%	77.58%	94.60%	80.30%	0.9730
ResNet34	89.81%	81.30%	79.46%	94.60%	80.80%	**0.9785**
ResNet50	89.81%	78.30%	79.04%	92.20%	78.30%	0.9782
EfficientNetB0	89.32%	83.20%	85.22%	91.50%	81.91%	0.9770
EfficientNetB2	92.23%	84.97%	80.96%	**96.10%**	82.03%	0.9662
ConvNeXt-B	**92.23%**	**89.10%**	**86.39%**	92.20%	**84.90%**	0.9770
ViT-B/16	90.29%	81.60%	81.79%	92.20%	81.50%	0.9690

While ConvNeXt-B evened out its other metrics with a precision of 89.10%, a recall of 86.39%, a specificity of 92.20%, and the best of them all, an F1 score of 84.90%, the test accuracy of both ConvNeXt-B and EfficientNetB2 was the same, 92.23%. ConvNeXt-B proved to be the best balance in all metrics, as it kept the sensitivity high without a significant rise in the number of false positives. The effectiveness of this model is further emphasized by the class-wise results. For Alopecia, the model obtained an accuracy of 98.54%, Recall of 100%, and F1 score of 96.10%, and for Dermatitis, the model obtained an accuracy of 99.51%, Recall of 100%, and F1 score of 98.36%, which is the best performance among all the classes. The model also had an excellent accuracy for Folliculitis with 100% precision and 96.77% F1 Score. Healthy Scalp achieved a more balanced performance with a 94.87% F1 score. Dandruff was comparatively worse, resulting in lower precision while keeping the recall high. The most difficult class was Hair Loss, with recall as low as 30.77% and an F1-score as low as 44.44%, despite a high accuracy and specificity. This

means that in the majority of cases, the model failed to predict actual Hair Loss samples more than it was predicting false ones.

The variations in performance between the tested models could be due to the different feature extraction methods and network architectures used by each model. DenseNet121 had a relatively high specificity of 94.60% and comparatively lower precision and recall, which could suggest that DenseNet121 was able to capture negative examples well, but not so well at distinguishing between visually similar scalp conditions. The overall accuracy was comparable for both ResNet34 and ResNet50, and the latter gave a slightly better F1-score and AUROC, suggesting that the shallow networks achieved enough representation capacity for the dataset without adding too much complexity. However, ResNet50 did not scale similarly, suggesting that the extra performance that the deeper (bottleneck) architecture offers does not translate to improvements for the relatively small and overlapping data set that is visual. EfficientNetB0 showed a good balance between precision and recall, whereas EfficientNetB2 showed the highest specificity of 96.10%, signifying that the compound scaling strategy may have helped this network extract discriminative features better and reduced the false-positive predictions, but only at the expense of missing out on some positive cases of the disease. The best overall performance was given by ConvNeXt-B, which achieved the highest accuracy, and also produced the highest precision and recall and F1 score, possibly because of its larger-kernel convolutions and updated feature extraction design [15] which enables it to learn fine-grained textural features as well as high-level patterns in scalp images. Lastly, the transformer-based model, ViT-B/16, performed competitively, but also had the lowest AUROC, potentially indicating that the relatively small dataset was not ideal for this model architecture, as Vision Transformers are typically more effective with extensive pretraining and training data [16]. The overall results indicate that the architectural design of each model had a remarkable impact on the sensitivity and specificity, and ConvNeXt-B achieved the best balance for the classification of six classes of scalp diseases. Table 5.2 shows a per-class wise result of the top performing model.

The ROC curve comparison in Figure 13 gives a threshold independent assessment of class-separability performance of all seven models. The models achieved a macro AUROC of 0.9730 for DenseNet121, 0.9785 for ResNet34, 0.9782 for ResNet50, 0.9770 for EfficientNetB0, 0.9662 for EfficientNetB2, 0.9770 for ConvNeXt-B, and 0.9690 for ViT-B/16. Overall, all models achieved AUROC values significantly greater than 0.5, highlighting the good performance of each of the evaluated architectures in discriminating between the six scalp condition categories. Although the overall classification metrics vary across the models, the relatively close grouping of the ROC curves shows that, overall, the models had broadly similar class-separability.

The ResNet34 model outperformed other models, with an AUROC of 0.9785, just slightly less than the score of ResNet50 (0.9782). This suggests that ResNet34 discriminated the best overall on different thresholds. The slight improvement of ResNet50 over ResNet34 in the margin suggests that adding more layers to the network was not a significant improvement in overall class separability for this dataset. This observation is similar to their overall accuracy scores (89.81%) whereas ResNet34 also achieved a slightly higher F1-score and AUROC. The result could suggest that there is not enough added representational power from ResNet50 to give a significant edge over the relatively sizeable data set of scalp images with similar visual content. Both EfficientNetB0 and ConvNeXt-B had an AUROC of 0.9770, which was comparable to the two ResNet models. They differed in their classification characteristics although they had the same AUROC. ConvNeXt-B

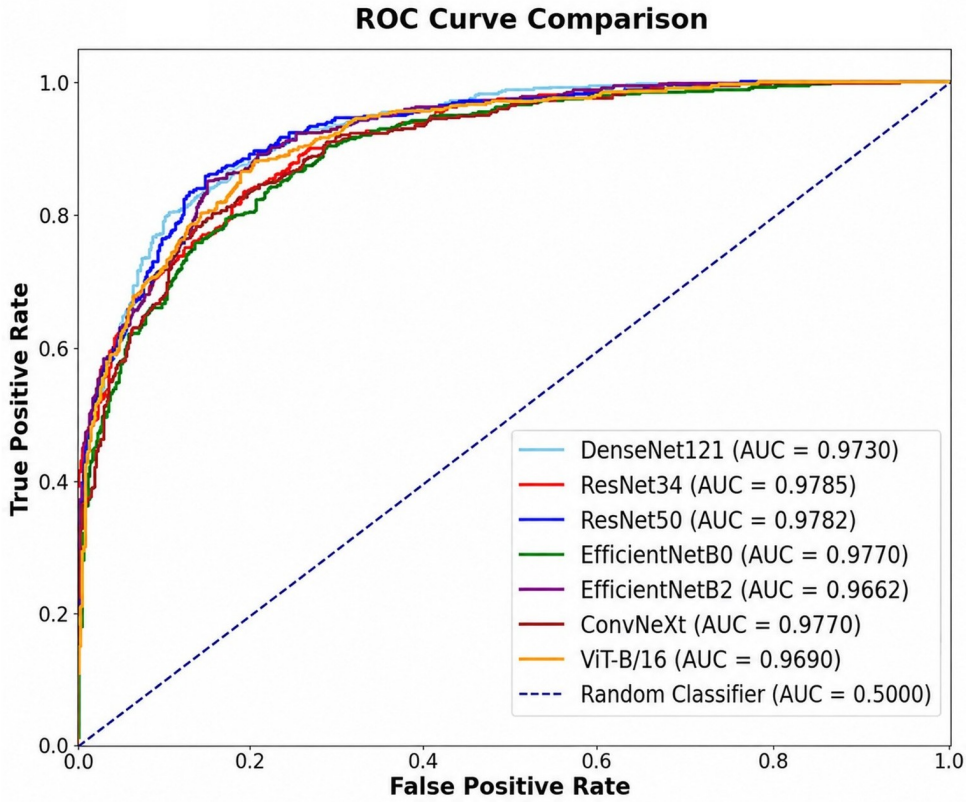

Figure 13: ROC curve comparison of the evaluated models.

achieved a much better recall and F1 score while EfficientNetB0 had a slightly lower overall accuracy. The comparable AUROC values show that both architectures were equally good at separating the classes at different thresholds, with a slightly better separation at the end for ConvNeXt-B. This difference makes it clear that AUROC is not necessarily the sole metric to consider when determining the best performing model; it measures discrimination over thresholds and not at a particular threshold that is used for making a final class assignment. DenseNet121's AUROC is 0.9730, just before the ConvNeXt-B model and the EfficientNetB0 model. In turn, the performance of ViT-B/16 was AUROC 0.9690, which is worse than CNN-based architectures, with the exception of EfficientNetB2. Nevertheless, its performance was still strong, but the AUROC was lower, indicating that the model had relatively more difficulties distinguishing some of the scalp disease categories. The model with the lowest macro AUROC (0.9662) was EfficientNetB2 which achieved the same highest AUROC (92.23%) as ConvNeXt-B. This is especially significant as it is not directly related to class separability, as the accuracy is high even though the class separability is not. EfficientNetB2 further obtained the greatest specificity of 96.10%, meaning that it was most successful in avoiding incorrect classifications. This conservative classification behaviour may, however, have led to a higher number of actual positive cases being missed as seen from its low recall of 80.96%. The lower AUROC of EfficientNetB2 suggests, however, that it was not as discriminative overall, as its performance was better, but not as good at the extremes of possible thresholds. The class-wise performance of the ConvNeXt-B model is presented in Table 5.2, providing a detailed evaluation of its classification performance across the six scalp condition categories.

The Dermatitis, Folliculitis, Alopecia, and Healthy Scalp showed the best overall performance with high accuracy, AUROC, precision, recall, and F1-scores for the class-wise results. Dermatitis was the highest performing class with an accuracy of 99.51%, AUROC

Table 5.2: Class-wise results of the top-performing model, ConvNeXt-B.

Class	Accuracy	Precision	Recall	F1-Score	AUROC
Alopecia	98.54%	92.50%	**100.00%**	96.10%	0.9950
Dandruff	96.12%	71.43%	88.24%	78.95%	0.9810
Dermatitis	**99.51%**	96.77%	**100.00%**	**98.36%**	**0.9990**
Folliculitis	99.03%	**100.00%**	93.75%	96.77%	0.9920
Hair Loss	95.15%	80.00%	30.77%	44.44%	0.9020
Healthy Scalp	96.12%	93.67%	96.10%	94.87%	0.9960

of 0.9990, and F1-score of 98.36%, whereas Folliculitis and Alopecia had a perfect precision and perfect recall respectively. The performance of dandruff was moderate, as it has less precision of 71.43%, but a higher recall of 88.24%. Hair Loss was by far the worst class with an AUROC of 0.9020, only 30.77% Recall and 44.44% F1 score. In general, it can be concluded that ConvNeXt-B has performed quite well on the classes having clearer visual separation, while Hair Loss has shown significant difficulty in classification, possibly because of its visual similarity to other scalp conditions.

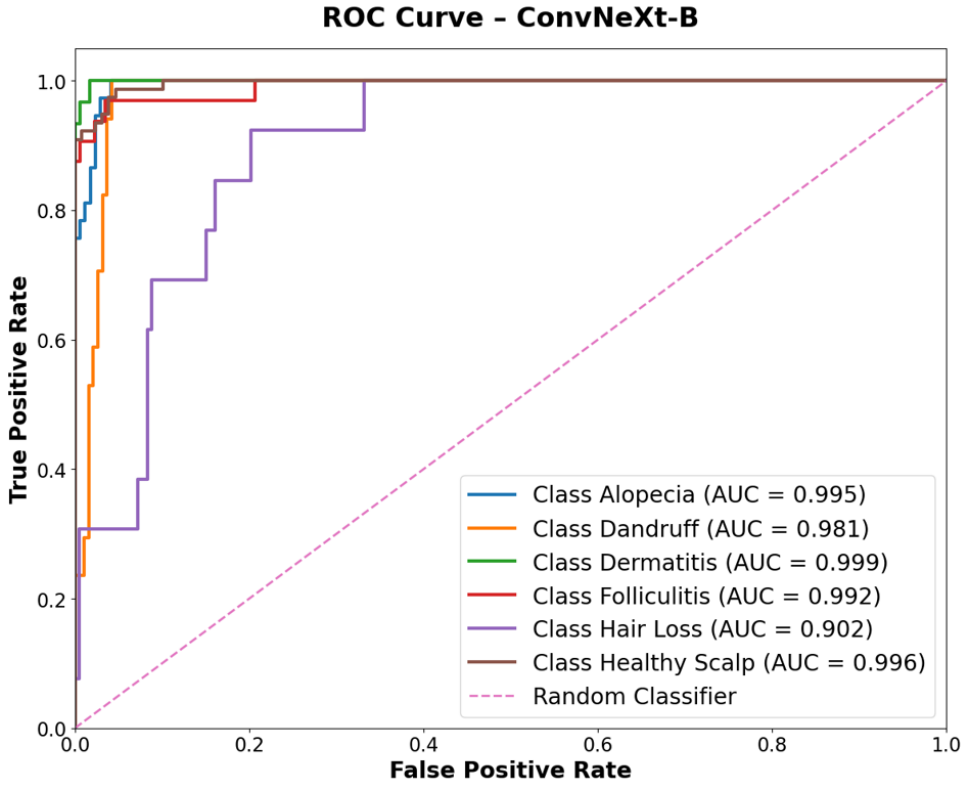

Figure 14: ROC class comparison of ConvNeXt-B.

The ConvNeXt-B model is further validated by the ROC analysis which shows that the model is able to classify the six condition categories of the scalp well. Macro AUROC of the model was 0.9770, which was high in the overall degree of separability of the classes. The highest class-wise AUROC was obtained for Dermatitis (0.999) followed by Healthy Scalp (0.996) and Alopecia (0.995) as shown in Figure 14. The near perfect AUROC obtained for the Dermatitis class indicates that the model performed well at discriminating between this class and the other scalp conditions at a number of classification thresholds. The result is also identical to the class-wise assessment where Dermatitis has an F1-score of 98.36% with a recall of 100.00% indicating that the model was able to capture the majority

of Dermatitis samples with high degree of accuracy. The strong AUROC values for the two classes of Healthy Scalp (98.37%) and Alopecia (96.10%) along with the F1-score of these two classes (94.87%, 96.10%) also show that the classes are well separated from the other classes. This is what might indicate that the visual properties of the classes in these classes were distinct enough for well-defined class-boundaries to be learnt from the model.

The AUROC of Folliculitis was 0.992 as well as excellent discriminative capability. While Dermatitis and Alopecia got the highest recall, the precision was 100% for the Folliculitis class and 96.77% for the F1-score, showing that the model was highly reliable in assigning an image to the Folliculitis class. Dandruff also exhibited a good AUROC value of 0.981 but performed comparatively less well compared to the other classes. This is in accordance with the lower precision (71.43%) and F1-score (78.95%) for the Dandruff class, which shows that there was a higher overlap in the prediction probabilities for this class. This could suggest that some of the Dandruff images have similarities to other scalp diseases and, therefore, pose a higher level of uncertainty in classification.

Similar to the class-wise metric analysis, as seen in the class-wise metric analysis, Hair Loss was the most difficult class for ConvNeXt-B with the lowest AUROC score of 0.902. The result of an AUROC > 0.90 is still considered as good discrimination, but the difference between Hair Loss and the other classes is high, suggesting it was harder for the model to discriminate well between Hair Loss and other scalp conditions. Additionally, its recall score of 30.77% and F1-score of 44.44%, which are significantly lower, further support this finding. The relatively low ROC performance indicates that the problem was not just with the classification threshold but the model's confidence scores for Hair Loss samples were generally less distinct from those of other classes. This could be because the symptoms of Hair Loss can be linked to other diseases, especially those that may also see less hair density or change in the visibility of the scalp. In summary, the ROC analysis shows that ConvNeXt-B achieved excellent discriminative performance in all six classes and an AUROC score greater than 0.90 for all classes. The results especially demonstrate the model's good performance in Dermatitis, Healthy Scalp, Alopecia, and Folliculitis classification and indicate that Hair Loss is a significant direction for improvement.

5.2 Augmented Dataset Results

Data augmentation resulted in performance alterations for all seven architectures. The performance of models was evaluated. Table 5.3 shows the results.

Table 5.3: Evaluation results of the models after augmentation.

Model Name	Accuracy	Precision	Recall	Specificity	F1-Score	AUROC
DenseNet121	89.32%	79.55%	80.68%	97.84%	79.26%	0.9687
ResNet34	89.32%	81.30%	82.70%	97.90%	81.70%	0.9620
ResNet50	90.78%	83.45%	84.53%	98.20%	83.68%	0.9780
EfficientNetB0	91.26%	85.36%	84.68%	98.25%	84.34%	0.9780
EfficientNetB2	**93.69%**	**88.73%**	**88.39%**	**98.70%**	**88.40%**	**0.9873**
ConvNeXt-B	92.23%	86.20%	87.20%	98.50%	86.60%	0.9760
ViT-B/16	90.78%	84.60%	86.00%	98.30%	85.00%	0.9810

From the results, EfficientNetB2 had the best accuracy (93.69%), macro precision (88.73%), macro recall (88.39%), macro specificity (98.70%) and F1-score (88.40%). EfficientNetB2 made improvements in all metrics compared to the results on the original dataset. This improvement in recall, with a very high specificity, indicates that the model has become better at identifying the disease cases whilst not necessarily increasing the number of false positive predictions. When considering the balance between sensitivity and specificity, EfficientNetB2 had the highest scores following augmentation. There was also a significant improvement in the class-wise results. The overall performance of most classes was good with Hair Loss being the most difficult. The class imbalance was a challenge, but augmentation was able to overcome some of these issues, however Hair Loss still found it a challenge as it had a small sample size and also looks similar to other classes.

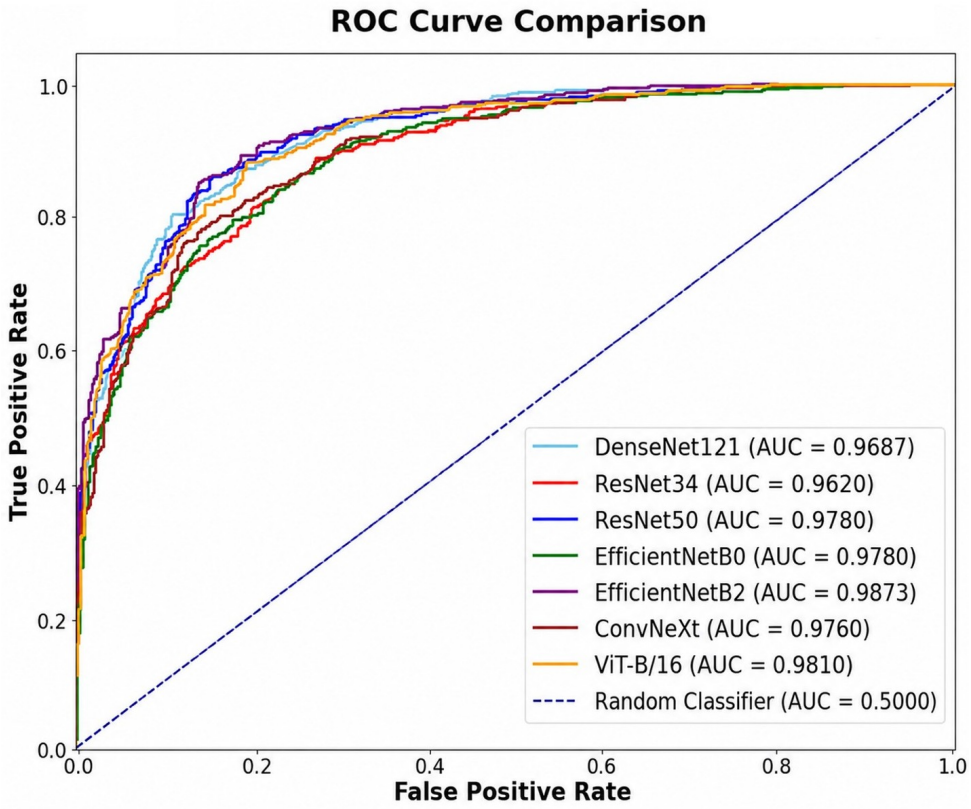

Figure 15: ROC curve comparison of the evaluated models with the augmented dataset.

EfficientNetB2 and ConvNeXt-B were the two methods that performed the best in the AUROC analysis, with macro AUROC values of 0.9901 and 0.9850, respectively. The difference between the two models was very small, which means that both architectures had excellent class separability. Their ROC curves were very similar and above the random-classifier baseline. Both models were able to discriminate well, with a marginally higher AUROC for EfficientNetB2, and both models showed significant improvements over the other architectures. The performance of DenseNet121 was the weakest overall for the augmented dataset. It was the least accurate of the two models, but was also the least precise, recall, and F1-score of all the models with the values 79.55%, 80.68%, and 79.26%, respectively, when combined with ResNet34. These reduced values did not affect its specificity, which was high at 97.84%, meaning that the model did not generate many false positives. DenseNet121 obtained a macro AUROC of 0.9687, demonstrating its good class separability despite having less classification accuracy when compared to the top-performing architectures. For the rest of the models, EfficientNetB0's accuracy

was 91.26% with an AUROC value of 0.9780, and ResNet50's accuracy was 90.78% with an AUROC of 0.9780. ViT-B/16 also gave an accuracy of 90.78%; however, it gave a slightly higher AUROC of 0.9810, which shows that the transformer-based architecture still held its own post-augmentation. The accuracy of ResNet34 was 89.32%, and it had the lowest AUROC of 0.9620 among the augmented models. The application of augmentation resulted in a mixed performance of all 7 architectures, especially the EfficientNetB2. In all the models, Hair Loss was always the most challenging class, likely due to visual similarity and the limited number of samples.

Table 5.4: Class-wise results of the top-performing model, EfficientNetB2.

Class	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUROC**
Alopecia	98.06%	92.31%	97.30%	94.74%	0.9930
Dandruff	97.09%	78.95%	88.24%	83.33%	0.9720
Dermatitis	**99.51%**	96.77%	**100.00%**	**98.36%**	**0.9980**
Folliculitis	99.51%	**100.00%**	96.88%	98.41%	1.0000
Hair Loss	95.15%	63.64%	53.85%	58.33%	0.8630
Healthy Scalp	98.06%	98.67%	96.10%	97.37%	0.9950

The Folliculitis class performed best with an AUROC of 1.0000, perfect accuracy (100.00%), and a very high F1 score (98.41%), which highlights its good class separability and very high accuracy in predicting the class. Dandruff, however, was the most poorly performing class, with the lowest AUROC score of 0.9720, precision of 78.95%, and F1-score of 83.33%, indicating more challenges in differentiating this condition from the other scalp conditions. Despite this, its recall of 88.24% indicates that the model was still able to identify most actual Dandruff cases. In general, the model achieved high accuracy and AUROC values for most of the classes, while the recall values were generally high for the different classes. Visual similarity between certain scalp conditions may have led to some false positive predictions, as shown by the lower precision for the Dandruff and Alopecia categories.

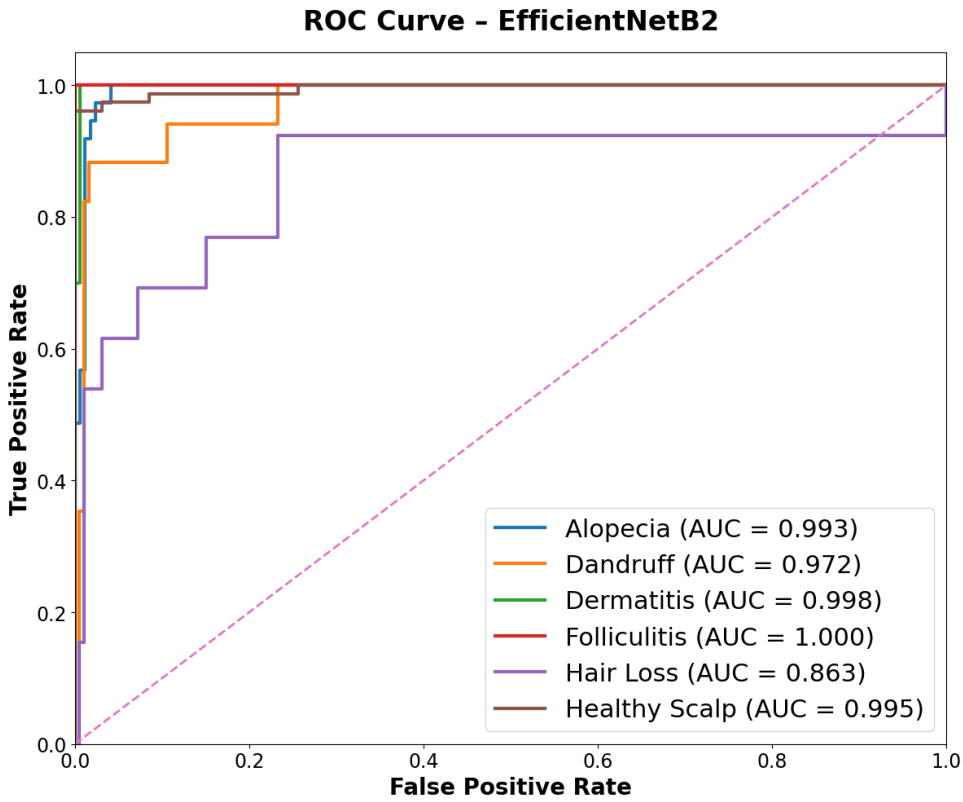

Figure 16: ROC class comparison of EfficientNetB2 on the augmented dataset.

The performance of the EfficientNetB2 model on the ROC analysis for each class is shown in Figure 16 and EfficientNetB2 has shown good class-separability performance across the six scalp condition classes. The greatest AUROC was for Folliculitis (1.000), followed by Dermatitis (0.998), Healthy Scalp (0.995) and Alopecia (0.993) showing excellent discrimination for all at all classification thresholds. Dandruff had the best AUROC equal to 0.972, but comparatively lower class separability than above mentioned classes. However, Hair Loss was the worst performing class for threshold independent discrimination, with the lowest AUROC of 0.863. In conclusion, the ROC analysis showed that EfficientNetB2 exhibited good class separability for Folliculitis, Dermatitis, Healthy Scalp and Alopecia and good discrimination for Dandruff and moderate discrimination for Hair Loss. However, all the six classes had AUROC values greater than 0.86, which suggests that EfficientNetB2 still had a significant discriminative power between each condition of the scalp and the other classes.

The study was deepened by tabulating the contrast between the augmented results and the baseline results using their non-absolute differences, called deviations. The mechanism is shown in Table 5.5.

Table 5.5: Comparison of deviations after augmentation (augmented result − baseline result).

Model Name	Accuracy Deviation	Precision Deviation	Recall Deviation	Specificity Deviation	F1-Score Deviation	AUROC Deviation
DenseNet121	−0.49%	−0.75%	+3.10%	+3.24%	−1.07%	−0.0043
ResNet34	−0.49%	0.00%	+3.24%	+3.30%	+0.90%	−0.0165
ResNet50	+0.97%	+5.15%	+5.49%	+6.00%	+5.38%	−0.0002
EfficientNetB0	+1.94%	+2.16%	+0.54%	+6.75%	+2.43%	+0.0010
EfficientNetB2	+1.46%	+3.76%	+7.43%	+2.60%	+6.37%	+0.0208
ConvNeXt-B	0.00%	−2.90%	+0.81%	+6.30%	+1.70%	+0.0008
ViT-B/16	+0.49%	+3.00%	+4.21%	+6.10%	+3.50%	+0.0120

The overall performance improvement of most of the evaluated models after data augmentation, which is shown in Table 5.3 when compared with the baseline results presented in Table 5.5 is clearly visible. The augmentation seems to have affected not just the accuracy of the overall classification, but also the accuracy of precision, recall, specificity, F1 score, and AUROC, with some architectures performing much better than others. Table 5.5 shows the deviations of each metric before and after augmentation. For each model the deviation magnitude is obtained by subtracting the baseline result metric from the augmented result metric. A positive (+ve) deviation indicates that the metric has improved following augmentation, a negative (-ve) deviation indicates that the metric has worsened following augmentation.

The baseline accuracy of 89.81% for DenseNet121 was marginally reduced by 0.49 percentage points, to 89.32% after augmentation. Its precision also decreased marginally from 80.30% to 79.55%, while recall improved from 77.58% to 80.68%. The specificity of the model significantly improved (from 94.60% to 97.84%), while the F1 score of the model slightly decreased (from 80.30% to 79.26%). Its AUROC also decreased from 0.9730 to 0.9687. The overall performance of DenseNet121 does not seem to be enhanced by augmentation, but perhaps did improve the recall and specificity, represented by the rise in overall positive and negative classification, respectively. These gains did not lead to better class separability or a more balanced mix of positive predictions, with a slight drop in precision and AUROC. This could be because DenseNet121's augmented data was not enriched with discriminative information that is useful to the model in a way that significantly enhances its overall feature representation. A similar trend was observed in the ResNet34 model pre and post augmentation, where accuracy dropped from 89.81% to 89.32%. But, its recall rose from 79.46% to 82.70%, and the specificity rose significantly from 94.60% to 97.90%. Its F1-score also improved from 80.80% to 81.70%. By comparison, its AUROC went down from 0.9785 to 0.9620, which was the biggest AUROC decline among the models that were tested. The results indicate that augmentation has had a positive impact on the final classification performance of the model at the selected threshold, even if it has had a negative impact on the class-separability performance of the model at the selected threshold. The higher recall and F1 value reflects a more balanced classification, while the lower AUROC suggests that the model's confidence scores were not as well balanced following augmentation, as indicated by the lower values across different thresholds.

ResNet50 had a more positive response to augmentation. Its accuracy increased from 89.81% to 90.78%, while precision improved from 78.30% to 83.45% and recall increased from 79.04% to 84.53%. Its specificity also increased from 92.20% to 98.20%, accompanied by an increase in F1-score from 78.30% to 83.68%. The AUROC stayed almost the same, marginally shifting between 0.9782 and 0.9780. This is a fairly stable AUROC and the improvements in nearly all other metrics prove that augmentation worked especially well for ResNet50. The model seems to have reached an improved compromise between detection of positive cases and prediction of false positives, without degrading the overall ability to separate the classes. The improvement could indicate that by introducing the extra image changes during augmentation, the deeper ResNet architecture learned more robust representations and without significantly changing its capacity to rank the samples across the classification thresholds [18].

The EfficientNetB0 model also saw significant gains after augmenting. Its accuracy increased from 89.32% to 91.26%, while precision improved from 83.20% to 85.36%. Recall decreased slightly from 85.22% to 84.68%, but specificity increased considerably from 91.50% to 98.25%. The F1-score improved from 81.91% to 84.34%, while AUROC increased from 0.9770 to 0.9780. The results suggest that augmentation helped the model to perform more optimally and significantly decreased the number of wrong positive predictions (as shown by the considerable increase in specificity). While the recall decreased slightly, the improvement in precision and F1 score indicates a shift in the model's reliability with positive predictions. The minimal rise in AUROC also suggests that augmentation could have been aiding EfficientNetB0 to learn slightly better class boundaries. The most striking results were seen for EfficientNetB2, which was the best performing model following augmentation. Its accuracy increased from 92.23% to 93.69%, while precision increased from 84.97% to 88.73% and recall increased from 80.96% to 88.39%. Specificity also increased from 96.10% to 98.70%, and the F1-score improved substantially from 82.03% to 88.40%. The AUROC was improved from 0.9662 to 0.9870, which was the greatest increase in AUROC compared to all the evaluated models. The results show that these augmentations had the significant benefit of EfficientNetB2 in almost all the evaluation metrics. The concurrent gains in precision, recall, specificity, and AUROC indicates that augmentation may have helped the model to learn more powerful and useful representation of features, leading to better classification performance and class separability. The improvement could also suggest that the extra scaling variations introduced were better suited to the compound scaling method used by EfficientNetB2, which enabled the model to learn more useful visual patterns while minimizing any sensitivity towards image variations in appearance [14].

In contrast, ConvNeXt-B exhibited mixed performance of augmentation. The baseline accuracy rate was 92.23% without any augmentation, and it continued as it is with augmentation, whereas precision rate decreased from 89.10% to 86.20% after augmentation. However, recall increased from 86.39% to 87.20%, and specificity increased from 92.20% to 98.50%. The F1-score increased from 84.90% to 86.60%, while AUROC increased considerably from 0.9770 to 0.9850. The lower precision shows the model made comparatively more false-positive predictions, but overall, the increases in recall, specificity, F1-score, and AUROC show that augmentation improved several aspects of the model's performance. Of particular interest, a significant gain in specificity indicates that the enhanced model was significantly more accurate in recognizing the negative cases. The AUROC value is also increasing, which means that there is better separability between the classes across all of the operating points, even though the precision for the chosen operating point is not

as high. Thus, the augmentation effect on ConvNeXt-B seems to be more complicated than that of EfficientNetB2, as it aided in improving various crucial metrics without compromising the overall accuracy. Lastly, the results of the ViT-B/16 clearly improved after augmentation. Its accuracy increased from 90.29% to 90.78%, precision increased from 81.60% to 84.60%, recall increased from 81.79% to 86.00%, and specificity increased from 92.20% to 98.30%. Its F1-score also increased from 81.50% to 85.00%, while AUROC improved from 0.9690 to 0.9810.

Overall, differences between the two tables demonstrate the impact of data augmentation on the different types of underlying model architectures. The overall accuracy and AUROC for DenseNet121 and ResNET34 decreased slightly, whereas the other models had increases in one or more of the main accuracy measures. The most significant overall improvements were found for the EfficientNetB2 model, with it achieving the highest accuracy after augmentation (93.69%) compared to 92.23% before augmentation. EfficientNetB0 and ResNet50 also showed consistently enhanced performance in several metrics, and ViT-B/16 consistently showed enhanced performance in the same metrics, while ConvNeXt-B maintained the same performance in terms of accuracy but enhanced in recall, specificity, F1 score, and AUROC. One major impact that was evident across the augmented results was an accentuated specificity of almost all the models, which may have helped the models to focus more on the target classes and become more effective in making fewer false positive predictions. The enhancements in recall seen for most models further point towards improved detection of positive cases, although this is not a consistent trend, as DenseNet121, EfficientNetB0 and EfficientNetB2 present varying changes in this metric. Overall, the results suggest that augmentation generally improved the robustness and generalization of the models tested, with EfficientNetB2 showing the biggest improvement, and the comparatively smaller or negative changes for DenseNet121 and ResNet34 indicate that the impact of augmentation might depend on how closely the image variations introduced by the augmentation align with the feature-learning abilities of the different models.

Chapter-6: Conclusion & Future Works

6.1 Summary

Scalp diseases are a large group of inflammatory, infectious, autoimmune and hair loss diseases affecting millions of people globally and can have a profound effect on physical health, psychological well-being and quality of life. Diagnosis is crucial for effective treatment, but many scalp conditions are visually similar and specialists are needed to make a diagnosis. With the fast development of deep learning, especially Convolutional Neural Networks (CNNs), the medical image analysis area has been revolutionized by automated extraction of complex visual features from images without relying on handcrafted feature engineering. Their learning of hierarchical representations has resulted in remarkable improvements in detection and classification of dermatological diseases such as scalp diseases. In more recent times, modern CNN architectures and transformer-based models have further improved the classification performance with better feature extraction, efficient network scaling and more representational capacity. These advancements have paved the way for the potential use of AI as a valuable decision-support tool in dermatology, offering the promise of enhancing diagnostic accuracy, supporting healthcare practitioners, and expanding access to quality screening, especially in settings with limited specialist healthcare resources.

6.2 Limitations

The results obtained from this thesis demonstrate the potential of automated classification of scalp diseases using state-of-the-art deep learning architectures, but there are some limitations that were encountered during the study. It is important to note these limitations as they represent the scope of the current study and indicate areas for future improvement.

One of the main drawbacks of this study is the lack of high-quality and clinically verified scalp disease image datasets. For this research, the dataset was gathered from Hair-Scalp-Analysis v2 dataset, which is publicly available. It offered good image samples for experimentation, but the dataset was not originally intended to be a fully balanced clinical benchmark. This meant that some classes had much more images than others and some disease classes had fewer samples. This imbalance in the classes impacted the learning process of the models particularly of the minority classes like Hair Loss and Dandruff. One drawback is that the data set had to be filtered and refined to be used for the final six-class classification problem. A number of original categories were dropped as they

were cosmetic hair conditions and not clinically relevant scalp diseases. Other labels were discarded as they were redundant severity stages rather than separate classes of disease. This cleaning process did make the dataset more appropriate for the thesis objective, but also decreased the number of images available for use. Hence, the final dataset size was kept relatively small as compared to large medical image datasets. The third limitation is related to class imbalance even after dataset refinement. There were the most samples in the Healthy Scalp class and the least samples in the Hair Loss class. This imbalance can lead to learning stronger representations of the majority classes and weaker representations of minority classes. The confusion matrix also revealed that the Hair Loss class was more challenging for the model to classify correctly compared to the other classes. This means that the overall accuracy does not necessarily reflect the accuracy for each disease category. Finally, the study was limited by computational resources and experimental scope. While seven modern deep learning architectures were compared in a standardized setting, the use of larger models, ensemble learning, explainable AI techniques, hyperparameter tuning, and clinical expert validation were not fully explored. These additions might enhance the reliability, interpretability, and usefulness of the system.

In summary, the major drawbacks of this study are the paucity of data, class imbalance, limited clinical validation, absence of external validation, lack of disease localization and no severity-level classification. These limitations aside, the study offers a systematic benchmark and a solid base for future research on AI-assisted scalp and hair disorder classification.

6.3 Analysis Of Social Effects

The proposed scalp-condition classification system can be beneficial in the social context by showing the application of deep learning in helping to identify common scalp conditions from images. Scalp issues can not only impact a person's health but also their confidence, appearance, and quality of life, especially when there are visible signs like hair loss, inflammation, or alterations in the scalp's condition. An automated image-based classification approach could help to make the preliminary assessment more accessible and consistent, especially in cases where dermatological expertise is scarce. The comparative experiments carried out in this study also highlight the need for robust models that can perform well even when the appearance of the images is different, as was done in this study before and after data augmentation. But the suggested system should be used as an auxiliary system and not as a substitute for skilled medical personnel. Misclassifications can lead to unwarranted anxiety or, on the other hand, to false reassurance; therefore, proper clinical validation and professional supervision are crucial before considering the use of such systems in the real world for diagnostics.

6.4 Environmental Impact Analysis And Sustainability

The environmental footprint of the proposed system is mostly related to the computational resources needed for image preprocessing, data augmentation, model training and evaluation. Training multiple deep-learning architectures, especially modern architectures

with many parameters, can require a lot of computational power and thus use electricity. The experimental comparison of models before and after augmentation also adds to the computational burden as augmented datasets might need more training iterations and processing. However, after training, an image-classification model can make individual predictions with significantly lower computational requirements than the training process. Hence, it is crucial to choose an architecture that offers a good compromise between prediction accuracy and the amount of computation needed to sustain deployment. Furthermore, data augmentation can increase the effective diversity of a relatively small dataset without having to collect and store an equal number of new images, thus saving the resources that are needed for large-scale additional data collection. Future implementations may be further optimized for sustainability by using efficient model architectures, optimized training procedures, transfer learning and energy-efficient computational infrastructure.

6.5 Ethical Issues

As the automated scalp-condition classification system evolves, there are various ethical issues to consider, including privacy of data, accuracy of diagnosis, bias, and proper usage. Medical and dermatological images can include sensitive data and so suitable precautions should be put in place to ensure that image data are collected, stored and processed in compliance with applicable privacy and data-protection principles. An additional factor to consider is whether the training data is representative of skin tone, hair features, image quality, age groups, and other demographic and clinical variations, which could lead to biased performance of the model. Because the model learns patterns from the training data it has, if the training data doesn't adequately represent a group, the model may perform less well for that group. In addition, the predictions obtained by the proposed system should not be considered as a medical diagnosis. Even a good model may give wrong positive and negative predictions as seen in the experimental results based on the class. Thus, any real deployment should involve suitable human oversight, model limitations transparency and clinical validation. The ultimate goal of the ethical use of such technology is to ensure that it is developed and used as a decision support tool and not a replacement for professional medical judgement.

6.6 Future Work

Future work would involve the collection of larger datasets of scalp images that were clinically verified. More diverse images would enhance the robustness of the model and minimize the risk of bias towards certain image conditions, camera settings, lighting environments, or patient groups. The current study employed a cleaned up version of an existing public dataset, so external validation on other datasets is necessary prior to any real-world deployment. Testing of the models on images gathered from various clinical or real-world sources would help better understand the extent to which the models can be generalized beyond the original images. Another important way of future research is to use 5-fold cross validation on the best models. ConvNeXt-B and EfficientNetB2 yielded the best overall performance in this study and could be further validated. Future experiments will be conducted on the base dataset only, and will be performed on 5-fold

cross validation for both ConvNeXt-B and EfficientNetB2. Here, the data set will be split into five different folds, with one fold being the test set and the other four folds being the training sets. This will give an idea of how consistent the models are in their performance across various train-test splits and not just one split. The CosineAnnealingLR scheduler can be used to enhance the training process for the five-fold cross-validation experiments. The CosineAnnealingLR will decrease the learning rate over the course of training according to a cosine curve, which means the model will be able to take larger steps at the beginning of training and smaller, more stable steps towards the end. This can facilitate smooth convergence of the models and enhance generalization performance. The method could be applied to ConvNeXt-B and EfficientNetB2 for a more solid comparison between the two best-performing architectures.

Possible techniques include horizontal flipping, rotation, random cropping, brightness adjustment, contrast adjustment, affine transformation, and synthetic sample generation. These techniques can be used to enhance the diversity of training and mitigate the impact of class imbalance particularly for underrepresented classes like Hair Loss and Dandruff. But augmentation should be carefully controlled to ensure that the medical meaning of the image is not distorted. Moreover, future study can be done on the base dataset only using ensemble learning techniques. ConvNeXt-B and EfficientNetB2 were found to have good performance but with different performance behavior, so it is possible to combine the prediction of both models to form a more reliable classification system. Ensemble approach can be done by averaging prediction probabilities, weighted averaging, or majority voting. This can enhance performance as one model might make up for the other's shortcomings. For instance, ConvNeXt-B can offer better overall performance, while EfficientNetB2 can offer better specificity. Both models, then, could be combined to provide a more stable and accurate experiment on the original base data set.

The best performance model could be used to create a real-time mobile application. The application might enable users to take scalp images and do on-device or cloud-based inference. The system could be used to screen for potential skin issues and suggest further evaluation by a dermatologist if any concerning skin changes are identified, rather than replacing the role of medical practitioners. This would render the research more practical and accessible. Lastly, explainable AI tools should be incorporated into the future versions of the system. Some methods like Grad-CAM or other visual explanation techniques could be used to visualize the areas of the image that affected the model's prediction. This would aid in assessing if the model is learning about the relevant scalp areas or background artefacts, hair lighting or irrelevant visual features. In medical image classification, explainability is particularly crucial as it enhances trust, aids in error analysis, and promotes transparency in AI-assisted diagnosis.

Bibliography

[1] B. E. Elewski, "Clinical diagnosis of common scalp disorders," in *Journal of Investigative Dermatology Symposium Proceedings*, vol. 10, no. 3. Elsevier, 2005, pp. 190–193.

[2] C. Ha, T. Go, and W. Choi, "Intelligent healthcare platform for diagnosis of scalp and hair disorders," *Applied Sciences*, vol. 14, no. 5, p. 1734, 2024.

[3] M. Du, Y. Wang, J. Wang, Y. Xie, Y. Xu, and D. Yang, "New frontiers of non-invasive detection in scalp and hair diseases: A review of the application of novel detection techniques," *Skin Research and Technology*, vol. 31, no. 2-5, p. e70163, 2025.

[4] Z. Yu, S. Kaizhi, H. Jianwen, Y. Guanyu, and W. Yonggang, "A deep learning-based approach toward differentiating scalp psoriasis and seborrheic dermatitis from dermoscopic images," *Frontiers in Medicine*, vol. 9, p. 965423, 2022.

[5] G. Rodríguez-Tamez, M. E. Herz-Ruelas, M. Gómez-Flores, J. Ocampo-Candiani, and S. Chavez-Alvarez, "Hair disorders in autoimmune diseases," *Skin Appendage Disorders*, vol. 9, no. 2, pp. 84–93, 2023.

[6] A. Crawl-Bey, J. Watts, J. Cotton, U. Nwaneri, S. Zinabu, M. Bisrat, E. Beyene, E. Pritchett, and M. Michael, "Beyond hair loss: Exploring the psychiatric burden of alopecia areata in a large cohort," *Journal of the National Medical Association*, 2025.

[7] D. N. N. Anh, P. N. Giau, B. N. Le Nguyen, A. Mai, N. Thi, M. Hien, and N. T. V. Tuyen, "Harnessing deep learning for scalp and hair disease classification: A comparative study of convolutional neural networks architectures," *Journal of Digital Information Management*, vol. 23, no. 2, 2025.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] W.-J. Chang, L.-B. Chen, M.-C. Chen, Y.-C. Chiu, and J.-Y. Lin, "Scalpeye: A deep learning-based scalp hair inspection and diagnosis system for scalp health," *Ieee Access*, vol. 8, pp. 134 826–134 837, 2020.

[10] M. Kim, Y. Gil, Y. Kim, and J. Kim, "Deep-learning-based scalp image analysis using limited data," *Electronics*, vol. 12, no. 6, p. 1380, 2023.

[11] M. Roy and A. T. Protity, "Hair and scalp disease detection using machine learning and image processing," *arXiv preprint arXiv:2301.00122*, 2022.

[12] L. J. Borda and T. C. Wikramanayake, "Seborrheic dermatitis and dandruff: a comprehensive review," *Journal of clinical and investigative dermatology*, vol. 3, no. 2, pp. 10–13 188, 2015.

[13] S. T. Bhuiyan, R. Rahman, R. Islam, M. Haque, M. Islam, S. Kobashi, and S. B. Alam, "Lung disease diagnosis from cxr using convolutional neural network (cnn): A comparative study," *Soft Computing*, vol. 29, no. 23, pp. 6333–6345, 2025.

[14] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PmLR, 2019, pp. 6105–6114.

[15] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2022, pp. 11 966–11 976.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. Ieee, 2017, pp. 2261–2269.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] mine, "scalp diseases dataset," https://universe.roboflow.com/mine-w0lar/scalp-diseases, may 2025, visited on 2026-08-19. [Online]. Available: https://universe.roboflow.com/mine-w0lar/scalp-diseases

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[21] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *2021 IEEE/CVF international conference on computer vision (ICCV)*. IEEE, 2021, pp. 32–42.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.