VITILIGO DETECTION USING MACHINE LEARNING ALGORITHMS

A Thesis Submitted
In Partial Fulfillment Of Requirements For The Degree Of

MASTERS OF TECHNOLOGY
in
DATA SCIENCE

by

Devansh Mathan
(2K23/DSC/27)

Under The Supervision Of

Prof. Ruchika Malhotra

(Head of Department, Department of Software Engineering)

Delhi Technological University

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2025

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

ACKNOWLEDGEMENT

We wish to express our sincerest gratitude to Prof. Ruchika Malhotra for her continuous guidance and mentorship that she provided us during the project. She showed us the path to achieve our targets by explaining all the tasks to be done and explained to us the importance of this project as well as its industrial relevance. She was always ready to help us and clear our doubts regarding any hurdles in this project. Without her constant support and motivation, this project would not have been successful.

Place: Delhi — Devansh Mathan

Date: 20.05.2025 — (2K23/DSC/27)

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

CANDIDATE'S DECLARATION

I, Devansh Mathan, 2K23/DSC/27 students of M.Tech (Data Science), hereby certify that the work which is being presented in the thesis entitled "Vitiligo Detection Using Machine Learning Algorithms" in partial fulfilment of the requirements for the award of degree of Master of Technology, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from Jan 2025 to May 2025 under the supervision of Prof. Ruchika Malhotra.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor (s)

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "Vitiligo Detection Using Machine Learning Algorithms" which is submitted by Devansh Mathan, Roll No – 2K23/DSC/27, Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Prof. Ruchika Malhotra
Head Of Department
Department of Software Engineering, DTU

Date: 20.05.2025

ABSTRACT

Vitiligo is a chronic skin condition involving the progressive depigmentation of the skin because of melanocyte destruction or malfunction. Prompt and proper diagnosis of vitiligo is critical, since early treatment can considerably enhance outcome and improve the quality of life for those suffering from the disease. Manual diagnosis, however, tends to be qualitative and time-consuming, especially in low-resource clinical settings. In this thesis, a diagnostic framework based on deep learning is presented for automatically diagnosing vitiligo from skin images. The research starts with the evaluation of five leading convolutional neural networks (CNNs), namely VGG16, ResNet50, InceptionV3, EfficientNet, and DenseNet121, on a publicly distributed vitiligo dataset retrieved from Kaggle.

The initial phase involved training and fine-tuning each CNN model to identify the top performers based on metrics such as accuracy, precision, recall, and F1-score. Among the evaluated models, VGG16, ResNet50, and DenseNet121emerged as the most effective, and were selected for further ensemble modeling. To enhance predictive reliability, three ensemble strategies were employed: bagging using Random Forest, boosting using XGBoost, and stacking with a logistic regression meta-learner. Beyond traditional ensemble methods, a Multilayer Perceptron (MLP)-based architecture was developed that fused deep features extracted from the three CNNs and learned complex inter-feature representations.

Experimental evaluations demonstrated that the proposed MLP-based model significantly outperformed all other approaches, achieving a classification accuracy of 99.22%, along with 99% precision, 99% recall, and 99% F1-score. Traditional ensembles such as Random Forest also performed well (98.43% accuracy), but were slightly less effective in terms of overall balance across evaluation metrics. These results confirm that feature-level fusion combined with neural modeling can yield superior classification outcomes in medical image analysis.

This research not only demonstrates the viability of deep ensemble learning for vitiligo detection but also sets the foundation for developing intelligent dermatological screening tools. The proposed framework is scalable and may be extended to support multi-class classification and other skin conditions in future clinical decision support systems.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Motivation

Vitiligo is a frequently occurring skin condition marked by patchy loss of pigmentation. It can either appear in small, localized areas or affect broader regions of the body. Across different global populations, its prevalence has been reported to range from 0.2% up to 1.8% in past researches[1][2]. Even though it is benign in nature, the toll of the disease on a patient is comparable to other diseases like seborrheic dermatitis and psoriasis[3]. Lesions on sites that are easily visible can trigger anxiety, shame and depression[4]. The enduring nature of vitiligo as well as the suboptimal treatments can considerably lower the well-being of the patient and might lead to deterioration of intrapsychic health[3-7].

While Diagnosis of vitiligo might not be too complicated, evaluating how severe or widespread the condition is can be difficult. The lesions don't follow a fixed pattern – can be of different size, shape or the location on the skin as well as the level of pigmentation loss. Also, when affected skin starts to repigment, can be due to treatment or naturally, it might not always come back evenly. This in turn can pose difficulty of tracking vitiligo[8][9]. Therefore, the motivation for this thesis to develop a reliable, AI-driven system that detects and analyses vitiligo lesions with remarkable accuracy. One that works across skin tones and doesn't rely on specialised tools. If executed properly, it could be a faster diagnosis with better monitoring for the people who need it.

1.2 Problem Statement

Even with rising adoption of deep learning in dermatology, vitiligo has still remained a less-studied condition. Unlike other pigmented skin lesions, vitiligo is a depigmented patch that can often blend subtly with surrounding skin. Also, current studies frequently evaluate using only a single model and rarely uses the ensemble techniques for robustness. Since single model CNN architectures often struggle with the generalisation due to dataset limitations, class imbalance and sensitivity to image variability.

Therefore, in order to fill these gaps, there is need for a comprehensive, comparative analysis of deep learning architectures for vitiligo detection, followed by the design of an ensemble feature based Neural classifier. Such a system can not only improve the detection accuracy but also provide a scalable real world clinical development.

1.3 Objectives

This thesis aims to contribute significantly to the field of skin disease detection by examine the various deep learning models on the vitiligo detection task. Specifically, it seeks to achieve the following objectives:

- To systematically evaluate and compare multiple deep convolutional neural networks—namely VGG16, ResNet50, InceptionV3, EfficientNet, and DenseNet121—on their ability to detect vitiligo from preprocessed dermoscopic and clinical skin images. This comparison is intended to identify the most suitable architectures for capturing depigmentation features critical to vitiligo diagnosis.
- To construct robust ensemble classification models by integrating the deep features of the top-performing CNNs. This involves implementing bagging (Random Forest), boosting (XGBoost), and stacking methods, as well as designing a custom Multilayer Perceptron (MLP) that fuses the learned representations into a single architecture. These approaches aim to enhance prediction consistency and leverage the complementary strengths of individual models.
- To critically analyze the diagnostic accuracy of each model based on clinically meaningful measures like accuracy, precision, recall, and F1-score, thus making sure that the system proposed is not only precise but sensitive and specific enough for practical use in dermatological screening.

1.4 Overview

The thesis is organised in the following sections: Chapter 1 highlights the motivation, problem statement and objectives behind the study. Chapter 2 discusses the research related to this field done till now. Chapter 3 elaborates the methodology in detail. The experimentation and results are analysed in Chapter 4. The thesis is finally completed with the concluding statements and future prospects in Chapter 5.

CHAPTER 2
RELATED WORK

Vitiligo, an acquired, chronic pigmentary skin disorder, is often misunderstood and misdiagnosed due to its similarity to other hypopigmented conditions. Affecting between 0.5% to 1% of the global population, vitiligo presents not only dermatological challenges but also profound psychological distress due to visible skin discoloration (Tanvir et al., 2024)[10]. Although it's not life-threatening, the disease often leads to anxiety, embarrassment, and in severe cases, social withdrawal. Conventionally, dermatologists rely on clinical inspection, often supported by Wood's lamp examination, to identify vitiligo. Yet this kind of diagnosis is still subjective and extremely variable depending on clinician experience, access to equipment, and lighting (Zhang et al., 2021) [11]. Particularly within primary healthcare or teledermatology arrangements, with limited availability of tools, diagnosis can be unreliable and imprecise.

Against this background, deep learning (DL) and machine learning (ML) methods have emerged as credible alternatives. These can result in data driven and replicable conclusions, minimizing reliance on human interpretation. Tanvir et al. (2024) [10] conducted a systematic literature review with the aid of computer assisted methods for the detection of vitiligo. They selected from an original list of 244 studies ten that were strict criteria-based—these included direct application of ML on skin images, provided quantitative accuracy measures, and did not include purely qualitative or non-image-based studies. Their results identified how ML can outperform or at least equal performance with dermatologists in certain diagnostic situations, most importantly distinguishing between vitiligo and look-alike disorders like pityriasis alba or tinea versicolor.

Additionally, Zhang et al. (2021) [11] compared the performance of convolutional neural networks (CNNs) like VGG-13, ResNet-18, and DenseNet-121 on both in-house and public datasets. They trained the models on thousands of clinical close-up images, and then tested them gains against the judgment of 14 human raters—dermatologists, residents, and general practitioners. Interestingly, the CNNs matched expert dermatologists. On the public dataset, DenseNet-121 resulted in an F1 measure of 0.9684, higher than the expert raters' 0.9221. These results highlight that with adequate training data, CNNs can provide diagnostic performance comparable to human experts—even without Wood's lamp imagery. The results also offer hope for rural or underserved regions lacking experienced dermatological professionals.

Still, classification alone is not enough. A more advanced approach was proposed by Guo et al. (2022) [12], who developed a hybrid deep learning model capable of both detecting and evaluating vitiligo lesions in terms of size and pigmentation. Their model comprised three stages: YOLOv3 for lesion detection, UNet++ for segmentation, and post-processing metrics (VAreaA, VAreaR, VColor) for morphometric and colorimetric analysis. Trained on a large dataset of DSLR-acquired images from Chinese patients with Fitzpatrick skin types III and IV, the model achieved 92.91%

sensitivity in lesion detection. For segmentation tasks, UNet++ attained a Jaccard Index of 0.79, outperforming other networks like UNet and PSPNet. Although the model performed slightly worse on lighter and darker skin tones (Fitzpatrick types I, II, and V), it still offered a reproducible and objective framework that matched dermatologists in precision and surpassed them in consistency.

However, training high-performance models requires abundant and diverse data. In medical imaging, especially dermatological datasets, image scarcity is a serious limitation. This is where generative models such as GANs come into play. Mondal et al. (2020) [13] addressed this issue by using Wasserstein GAN with Gradient Penalty (WGAN-GP) to generate synthetic skin lesion images—including vitiligo. Their pipeline involved three main steps: dataset preprocessing (including contrast normalization and morphological filtering), synthetic sample generation using WGAN-GP, and classification using CNNs. After augmenting the dataset with 1,504 synthetic images, the model (DenseNet-121) reached a classification accuracy of 94.25%, nearly 11% higher than the non-augmented baseline. These results emphasize how data augmentation via GANs can significantly improve DL model performance, especially when dealing with underrepresented skin conditions like vitiligo.

Despite the optimism, there are limitations. One concern repeatedly mentioned in the literature is the lack of standardization across datasets. Many studies focus on single-race populations or are limited to specific skin types, reducing the generalizability of results. For example, the hybrid AI model by Guo et al. (2022) [12] worked best on Fitzpatrick types III–IV but underperformed on types I and V. Another issue is interpretability. While CNNs can classify and segment with high accuracy, the black-box nature of these models still raises skepticism among clinicians. Without clear visual or statistical explanation, many are reluctant to adopt AI-driven tools into their workflow. Moreover, as Tanvir et al. (2024) [10] rightly argue, several studies still report insufficient methodological rigor—such as small sample sizes, poor validation techniques, or lack of follow-up studies—making the results difficult to replicate or deploy in clinical settings.

Nonetheless, these limitations open the door for further research. Larger multi-ethnic datasets, explainable AI frameworks, and integration of clinical metadata (e.g., patient history, lesion progression) can help build robust diagnostic pipelines. Ensemble techniques that combine CNN outputs with dermatological scoring systems like VASI or VETF might also improve interpretability and clinical trust. And perhaps most importantly, all models must be tested not only on benchmark datasets but in real-world clinical scenarios to ensure usability.

To sum up, deep learning models are showing remarkable promise in vitiligo detection and analysis. CNN-based architectures like DenseNet, UNet++, and YOLOv3 offer superior classification and segmentation capabilities. When augmented with GAN-generated synthetic data, these models overcome the challenge of limited training samples. Hybrid systems, as developed by Guo et al. (2022) [12], demonstrate that morphometric and colorimetric lesion analysis can be fully automated, supporting both research and clinical applications. Still, to make AI an integral part of dermatological diagnostics, future studies must address gaps in dataset diversity, model explainability, and real-world validation.

Vitiligo, an acquired, chronic pigmentary skin disorder, is often misunderstood and misdiagnosed due to its similarity to other hypopigmented conditions. Affecting between 0.5% to 1% of the global population, vitiligo presents not only dermatological challenges but also profound psychological distress due to visible skin discoloration (Tanvir et al., 2024) [10]. Although it's not life-threatening, the disease often leads to anxiety, embarrassment, and in severe cases, social withdrawal. Conventionally, dermatologists rely on clinical inspection, often supported by Wood's lamp examination, to identify vitiligo. However, this diagnostic approach remains subjective and varies widely based on clinician experience, equipment access, and lighting conditions (Zhang et al., 2021) [11]. Especially in primary healthcare or teledermatology setups, where access to tools is minimal, diagnosis can be both inconsistent and inaccurate.

Against this background, deep learning (DL) and machine learning (ML) methods have emerged as credible alternatives. These can result in data driven and replicable conclusions, minimizing reliance on human interpretation. Tanvir et al. (2024) [10] conducted a systematic literature review with the aid of computer assisted methods for the detection of vitiligo. From an initial pool of 244 studies, they shortlisted ten that met strict criteria—these involved direct use of ML on skin images, reported quantitative accuracy metrics, and excluded purely qualitative or non-image-based research. Their findings highlighted how ML could outperform or at least match dermatologists in some diagnostic settings, particularly in distinguishing vitiligo from look-alike conditions such as pityriasis alba or tinea versicolor.

Further, Zhang et al. (2021) [11] evaluated the performance of convolutional neural networks (CNNs) such as VGG-13, ResNet-18, and DenseNet-121 across both in-house and public datasets. They trained these models on thousands of clinical close-up images, then tested them against the judgment of 14 human raters—including dermatologists, residents, and general practitioners. Remarkably, the CNNs performed comparably to expert dermatologists. On the public dataset, DenseNet-121 yielded an F1 score of 0.9684, exceeding the expert raters' 0.9221. These findings underscore that with proper training data, CNNs can achieve diagnostic accuracy that rivals human experts—even in the absence of Wood's lamp imagery. The results also offer hope for rural or underserved regions lacking experienced dermatological professionals.

Still, classification alone is not enough. A more advanced approach was proposed by Guo et al. (2022) [12], who developed a hybrid deep learning model capable of both detecting and evaluating vitiligo lesions in terms of size and pigmentation. Their model comprised three stages: YOLOv3 for lesion detection, UNet++ for segmentation, and post-processing metrics (VAreaA, VAreaR, VColor) for morphometric and colorimetric analysis. Trained on a large dataset of DSLR-acquired images from Chinese patients with Fitzpatrick skin types III and IV, the model achieved 92.91% sensitivity in lesion detection. For segmentation tasks, UNet++ attained a Jaccard Index of 0.79, outperforming other networks like UNet and PSPNet. Although the model performed slightly worse on lighter and darker skin tones (Fitzpatrick types I, II, and V), it still offered a reproducible and objective framework that matched dermatologists in precision and surpassed them in consistency.

However, training high-performance models requires abundant and diverse data. In medical imaging, especially dermatological datasets, image scarcity is a serious limitation. This is where generative models such as GANs come into play. Mondal et al. (2020) [13] addressed this issue by using Wasserstein GAN with Gradient Penalty (WGAN-GP) to generate synthetic skin lesion images—including vitiligo. Their pipeline involved three main steps: dataset preprocessing (including contrast normalization and morphological filtering), synthetic sample generation using WGAN-GP, and classification using CNNs. After augmenting the dataset with 1,504 synthetic images, the model (DenseNet-121) reached a classification accuracy of 94.25%, nearly 11% higher than the non-augmented baseline. These results emphasize how data augmentation via GANs can significantly improve DL model performance, especially when dealing with underrepresented skin conditions like vitiligo.

Despite the optimism, there are limitations. One concern repeatedly mentioned in the literature is the lack of standardization across datasets. Many studies focus on single-race populations or are limited to specific skin types, reducing the generalizability of results. For example, the hybrid AI model by Guo et al. (2022) [12] worked best on Fitzpatrick types III–IV but underperformed on types I and V. Another issue is interpretability. While CNNs can classify and segment with high accuracy, the black-box nature of these models still raises skepticism among clinicians. Without clear visual or statistical explanation, many are reluctant to adopt AI-driven tools into their workflow. Moreover, as Tanvir et al. (2024) [10] rightly argue, several studies still report insufficient methodological rigor—such as small sample sizes, poor validation techniques, or lack of follow-up studies—making the results difficult to replicate or deploy in clinical settings.

Nonetheless, these limitations open the door for further research. Larger multi-ethnic datasets, explainable AI frameworks, and integration of clinical metadata (e.g., patient history, lesion progression) can help build robust diagnostic pipelines. Ensemble techniques that combine CNN outputs with dermatological scoring systems like VASI or VETF might also improve interpretability and clinical trust. And perhaps most importantly, all models must be tested not only on benchmark datasets but in real-world clinical scenarios to ensure usability.

To sum up, deep learning models are showing remarkable promise in vitiligo detection and analysis. CNN-based architectures like DenseNet, UNet++, and YOLOv3 offer superior classification and segmentation capabilities. When augmented with GAN-generated synthetic data, these models overcome the challenge of limited training samples. Hybrid models, used by Guo et al. (2022) [12], show the potential of morphometric and colorimetric lesion analysis to be fully automated, paving the way for research and clinical applications. For AI to become a standard tool in dermatological diagnosis, however, the current research must bridge gaps in dataset diversity, model interpretability, and practical utility.

Previous work has indicated additional classification accuracy and interpretability enhancements with modern architectures. Zhong et al. (2024) [14] investigated the effectiveness of ResNet and Swin Transformer models in the diagnosis of vitiligo. Their paper published in Scientific Reports was targeted not only at performance measures but also towards enhancing model prediction interpretability.The application of the marriage of convolutional and transformer-based methods by researchers yielded encouraging results, vindicating hybrid deep learning's contribution models to dermatology.

In another initiative, Kantoria et al. (2020) [15] investigated CNN-based classification models particularly for vitiligo with less complicated methodology. Having been published by the International Research Journal of Engineering and Technology (IRJET), their research explained how CNNs, even without being highly pretrained or augmented, could still perform good classification in accuracy when used with domain-specific tuning. Being less complicated than transformer models, their method is still worth it for quick deployment in resource-constrained environments.

Another work by Bashar and Suliman (2022) [16] examined whether pre-trained CNN architectures can be used to classify vitiligo images. Their thesis too explored the economic implications of automating vitiligo detection at clinical levels. The study indicated how pretrained networks like VGG and ResNet can reduce diagnostic costs and optimize healthcare workflows without compromising diagnostic reliability.

Similarly, Thanka et al. (2020) [17] proposed a multi-class skin disease classification system using deep CNNs. Although their study wasn't focused solely on vitiligo, the framework included vitiligo as one of the disease classes. Published in the Journal of Green Engineering, the research showed high accuracy across multiple dermatological conditions and demonstrated how CNN-based approaches can be adapted for broad-spectrum skin disease diagnosis, including vitiligo.

Table 1 : Literature survey summary table

S. No.	Paper Refereneces	Journal/Conference	Year	Performance Reported
1	[12]	Annals of Translational Medicine	2022	YOLOv3 detection sensitivity: 92.91%, UNet++ segmentation JI: 0.79
2	[11]	Frontiers in Medicine	2021	DenseNet F1 Score: 0.96 (public dataset); Outperformed expert raters
3	[10]	BioMed Research International	2024	Review paper – highlighted ML potential and gaps, no specific

				performance metric
4	[13]	IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)	2020	DenseNet-121 accuracy: 94.25% with GAN augmentation
5	[14]	Scientific Reports	2024	Reported enhanced accuracy and interpretability using ResNet and Swin Transformer; Accuracy: 94%
6	[15]	International Research Journal of Engineering and Technology (IRJET)	2020	CNN-based vitiligo classification; performance validated through accuracy comparisons; Accuracy: 96.5%
7	[16]	M.S. Thesis, KTH Royal Institute of Technology	2022	ResNet Accuracy: 85.6%; InceptionV3 Accuracy : 91%
8	[17]	Journal of Green Engineering	2020	DCNN Accuracy: 96%

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Framework Overview

The proposed framework for vitiligo detection is designed as a multi-stage pipeline as shown in Figure 3.1 that integrates deep learning and ensemble methods to enhance classification performance. It begins with acqusition of skin images from the Kaggle Vitiligo dataset, followed by a series of preprocessing steps including resizing, normalisation. Five pre-trained CNN models – VGG16. ResNet50, InceptionV3, EfficientNet, DenseNet121 – are finetuned on the dataset to assess their individual classification accuracy. Based on their performance, the top three models (VGG16. ResNet50, DenseNet121) are selected for futher ensemble construction. These selected models are then used to generate deep features which are fed into three enseble strategies: bagging using Random Forest, boosting with XGBoost and stacking. In addition to these, a custom ensemble is developed where the deep features from three CNNs are concentrated and passed through a Multilayer Perceptron network, which ultimately achieved the highest accuracy and f1 score. The framework ensures that each phase – from preprocessing to ensemble fusion – is optimized to deal with the challenges of class imbalance.

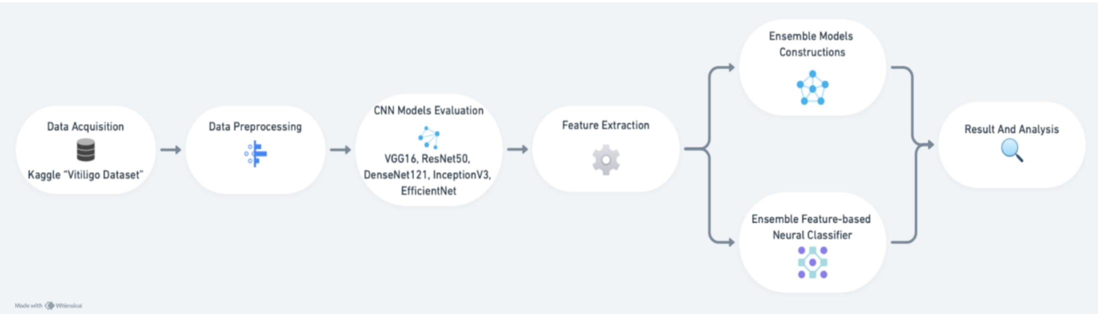


Figure 3.1 : Architecture of proposed methodology

3.2 Dataset Profile

The dataset used in this thesis is obtained from Kaggle, Titled *”Vitiligo Dataset”,* which consits of labled skin images divided into two categories: ‘Healthy Skin’ and ‘Vitiligo’. This dataset contains a total of 1271 RGB images. 891 images are labeled as healthy skin and 380 labled as vitiligo. Each Each image is a real-world clinical or dermoscopic capture, varying in resolution, background noise, lighting conditions, and skin tones, providing a diverse and realistic set of examples. Figure 3.2 shows sample image from the dataset having two classes in which one is healthy skin and another is vitiligo.

Figure 3.2 Sample Image from the dataset having two classes i.e. Healthy and Vitiligo

To guarantee the model generalizability and balanced learning, the data was stratified and divided into three subsets:

- Training Set (70%) – employed for model training and parameter tuning.
- Validation Set (10%) – utilized for hyperparameter tuning and avoiding overfitting.
- Test Set (20%) – used for performance benchmarking and model final assessment.

Due to imbalance of class—fewer vitiligo images compared to normal images—data augmentation techniques such as rotation, horizontal flip, zoom, and brightness adjustments were applied during training. All the images were resized to 224×224 pixels.

3.3 Data Preprocessing

Effective preprocessing is essential in medical image analysis, especially for conditions like vitiligo where lesion visibility is often subtle and image quality can vary. To prepare the dataset for deep learning models and ensure consistency across the input pipeline, several preprocessing steps were applied.

All input images were first resized to 224×224 pixels, which is the standard input dimension required by most pre-trained CNN architectures such as VGG16, ResNet50, and DenseNet121. This resizing step ensures compatibility and reduces computational overhead while preserving essential visual features.

Next, to improve the local contrast of skin regions, especially the depigmented patches characteristic of vitiligo, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied. CLAHE enhances the contrast of each image in a localized manner, allowing better distinction between affected and unaffected skin areas without over-amplifying noise.

Following enhancement, all pixel values were normalized to a range of [0, 1] by dividing the RGB values by 255. This normalization helps stabilize the training process and ensures faster convergence of the neural networks.

To address the issue of class imbalance and enhance the model's robustness, data augmentation techniques were applied exclusively to the training set. These augmentations introduced variability in the input images, helping the model generalize better to unseen data. Specifically, the images were randomly rotated up to ±20 degrees, flipped both horizontally and vertically, and subjected to zooming and spatial shifting to simulate different viewing perspectives. Additionally, brightness and contrast jittering were incorporated to mimic variations in lighting conditions. These transformations expanded the diversity of the training data without altering the semantic content of the images, thereby improving the network's ability to detect vitiligo under varied clinical scenarios.

Finally, all images were converted into NumPy arrays and reshaped into the appropriate tensor formats required for model training. This standardized preprocessing pipeline ensured that both deep learning and ensemble models received high-quality, uniformly formatted input for optimal learning and evaluation.

3.4 Base CNN Models Evaluation

In an effort to develop a reliable ground for vitiligo classification, five state-of-the-art convolutional neural network (CNN) models were chosen and evaluated VGG16, ResNet50, InceptionV3, EfficientNet, and DenseNet121. The models were chosen because they had already demonstrated success in other object or image inspection task, from medical imagery to dermatological assessment. All the models offer unique architectural features, hence their applicability to comparative performance analysis when used in the scope of vitiligo detection.

All models were pre-trained with ImageNet-pretrained weights to leverage transfer learning. The last classification layers were substituted with special fully connected layers appropriate for binary classification (Healthy vs. Vitiligo). For maintaining uniformity, the same input size (224×224×3) was fed to each model, and they were trained on the same preprocessed dataset with the same training-validation splits.

Training was performed for 20 epochs, subject to early stopping behavior tracked via validation loss. The Adam optimizer with a low learning rate was utilized for fine-tuning convolutional layers without destabilizing pretrained features. Binary cross-entropy loss was selected as the objective function because the task of classification is binary.

For equitable comparison, the performance was evaluated with the same test set and standard parameters like accuracy, precision, recall, and F1-score. These values enabled the quantification of not only the overall accuracy of predictions but also the sensitivity and specificity of the model towards identifying vitiligo lesions, which is highly relevant in medical diagnosis.

The outcomes indicated that the five models all worked fairly well, but VGG16, ResNet50, and DenseNet121continuously surpassed InceptionV3 and EfficientNet on all the measures.

3.5 Feature Extraction

Following the evaluation of individual CNN models, the top three performing architectures—VGG16, ResNet50, and DenseNet121—were selected for feature extraction. Rather than using their final softmax predictions directly, the proposed framework leverages the deep feature representations generated by these models to serve as input for subsequent ensemble learning.

Each of the selected CNNs was modified such that the classification head was removed. Instead, features were extracted from the penultimate layer, typically the global average pooling layer, which captures high-level semantic information while preserving spatial invariance. This layer outputs a fixed-length feature vector for each image that encapsulates the learned representations from the convolutional filters.For every image in the dataset, the corresponding deep features were extracted independently using each of the three CNNs. This resulted in three separate feature vectors per image, each corresponding to a different model. These vectors were then concatenated to form a comprehensive feature representation combining insights from all three networks. This fusion of multi-model features was designed to exploit the diversity in the internal representations of the models—VGG16's structured simplicity, ResNet50's residual connections, and DenseNet121's dense block connectivity.

This approach also reduces computational complexity in the ensemble stage, since the heavy lifting of feature extraction is handled by the pre-trained CNNs, and the downstream classifiers deal only with fixed-size vector representations. Additionally, the extracted features retain model-specific patterns which, when fused, offer a more robust and generalizable representation of vitiligo lesions across varying skin tones, lighting conditions, and image qualities.

3.6 Ensemble Model Construction

In order to further improve the robustness and precision of vitiligo classification, the research adopts several ensemble learning methods based on the deep features extracted from the top-performing three CNN models—VGG16, ResNet50, and DenseNet121. Ensemble methods are also known to enhance generalization by combining the strength of more than one base model and reducing their respective weaknesses. Three popular adopted ensemble methods were tested in this study: bagging, boosting, and stacking.

3.6.1 Bagging: Random Forest Classifier

The bagging approach, involves training several weak learners on random subsets of the dataset and aggregating their predictions. In this case, the Random Forest classifier was used, which constructs several decision trees on different bootstrap samples of the data. Each tree receives the same concatenated deep feature vectors (from VGG16, ResNet50, and DenseNet121) as input. The final prediction is obtained through majority voting. Random Forest is especially good at reducing overfitting and works well with high-dimensional data, so it is a good choice for feature-level ensemble learning.

3.6.2 Boosting: XGBoost Classifier

The boosting approach, applied through XGBoost (Extreme Gradient Boosting), trains sequentially a chain of models where each model attempts to fix the mistakes committed by the previous one. The input to XGBoost is once more the merged feature set of the three CNNs. While bagging is primarily centered on easy-to-classify instances, boosting concentrates relatively more on difficult-to-classify samples and generally leads to improved bias reduction. In this analysis, XGBoost performed excellently because of its capability to deal with complicated feature interactions, and the in-built regularization prevented overfitting.

3.6.3 Stacking Ensemble

In the stacking ensemble, the predictions (probability scores) from the individual CNN models are treated as inputs to a meta-classifier, which learns to combine them into a final prediction. For this task, a logistic regression model was used as the meta-learner. Each base model (VGG16, ResNet50, and DenseNet121) was trained independently, and their outputs on the validation set were used to train the meta-classifier. The idea behind stacking is to allow the meta-learner to capture patterns and correlations in the predictions that individual models might miss.

3.7 MLP – Based Feature Fusion Model

The MLP architecture consists of several key components designed to effectively learn from the concatenated feature vectors derived from the top-performing CNN models. The input layer of the network receives a high-dimensional feature vector formed by combining the outputs from the global average pooling layers of VGG16, ResNet50, and DenseNet121. This fused vector carries rich semantic information from each model, capturing diverse patterns in the input images.

Following the input, the architecture includes two or more hidden layers, each composed of dense (fully connected) neurons. These layers employ the ReLU activation function to introduce non-linearity, which allows the network to model complex interactions among features. The number of neurons in each layer was determined empirically to maintain a balance between computational efficiency and learning capacity. To address overfitting, dropout layers were incorporated between the dense layers. During training, dropout randomly disables a fraction of the neurons, which encourages the network to generalize better by not relying too heavily on specific nodes. The final layer of the model is a single neuron activated by a sigmoid function, producing a probability value indicating whether the input image is classified as vitiligo or healthy skin.

For training, the MLP was optimized using the Adam optimizer, known for its fast convergence and adaptive learning rates. A binary cross-entropy loss function was employed since the task involves binary classification. A small learning rate was selected to ensure stable fine-tuning of the network weights. The training process incorporated early stopping, where the validation loss was continuously monitored, and training was halted if no improvement was observed over several epochs. This helped prevent overfitting and ensured the model retained the best weights.

Algorithm 3.1: Vitiligo Detection Framework Based on CNN Feature Fusion and Ensemble Learning

Given:

- A labeled dataset $D = (x_i,\ y_i)_{i=1}^{N}$, where $x_i \in R^{H\times W\times C}$ are color skin images and $y_i \in \{0,1\}$ indicates the class (0 = healthy, 1 = vitiligo).
- Pre-trained CNN models $M_1, M_2, \ldots, M_5 \in$ *{VGG16, ResNet50, InceptionV3, EfficientNet, DenseNet121}*.

Output:

- Classification function f:$R^{H\times W\times C} \rightarrow \{0,1\}$

Step 1: Preprocessing

1.1 **Resize:**

Each image $x_i \rightarrow x_i' \ \in \ R^{224\times 224\times 3}$

1.2 **CLAHE (Contrast Enhancement):**

Apply CLAHE transformation: $x_i'' = CLAHE(x_i')$

1.3 **Normalization:**

Rescale each pixel value to [0,1]: $x_i''' = \frac{x_i''}{255}$

1.4 **Augmentation:**

Define augmentation transformations T=*{rotate,flip,zoom,jitter}* and apply to x_i''' for synthetic sample generation.

Step 2: CNN Model Training and Evaluation

Train each model M_k on the training set $D_{train} \subset D$, and evaluate using:

$Accuracy\ =\ \frac{TP+TN}{TP+TN+FP+FN}$, Precsion= $\frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN} F1 - score = \frac{2 \cdot \Pr ecision \cdot Recall}{\Pr ecision + Recall}$

Select top 3 models:

$$S = \{M_{VGG16}, M_{ResNet50}, M_{DenseNet121}\}$$

Step 3: Feature Extraction

From each selected CNN model $M_k \in S$, extract feature vectors from global average pooling layer:

$$f_i(k) = M_k GAP(x_i''') \ \in \ R^{d_k}$$

Concatenate features:

$$F_i = [f_i(1) \parallel f_i(2) \parallel f_i(3)] \in R^d,\ d = d_1 + d_2 + d_3$$

Step 4: Ensemble Classifier Construction

Using Fi as input features, define classifiers:

- Bagging: $f_{RF}(F_i) \rightarrow y_i$
- Boosting: $f_{XGB}(F_i) \rightarrow y_i$
- Stacking: $f_{stack}(F_i) = g\big(f_1(F_i), f_2(F_i), f_3(F_i)\big)$

Train all classifiers on $F_{train} = (F_i, y_i)$

Step 5: Custom MLP Feature Fusion

Define MLP as a function $f_{MLP}: R^d \rightarrow [0,1]$:

$$h_1 = ReLU(W_1 F_i + b_1),\ h_2 = ReLU(W_2 h_2 + b_2)\ y_i = \sigma(W_3 h_2 + b_3)$$

Where σ is the sigmoid function. Optimize using binary cross-entropy:

$$L = -[y_i \log(y_i) + (1 - y_i) \log(1 - y_i)]$$

Step 6: Model Evaluation

Compare all models f_{CNN}, f_{RF}, f_{XGB}, f_{stack}, f_{MLP} on the held-out test set using classification metrics.

$$f_{best} = \arg\max_{\{f} \quad F1 - score(f)$$

3.8 Model Evaluation Metrics

For Objective evaluation and comparison of the performance of all models – individual CNNs, ensemble classifiers, and the proprietary MLP architecture – four broadly used evaluation measures were employed: accuracy, f1-score, recall, precision.

Accuracy quantifies the overall accuracy of the model by measuring the ratio of correctly predicted samples to total samples. Although useful, it ma be deceptive in imbalanced datasets. Therefore, precision and recall were also taken into account. Precision measures the ratio of true positive predictions to total positive predictions, and this is important for minimising false positives. Recall , by contrast, tracks how well the model accurately identifies all positive cases that are actual, and this is critical in medicine where a missed case (false negative) may have severe consequences.

For balancing recall and precision, the F1-score was utilized, which is the harmonic mean of the two. It offers one metric that takes both false positives and false negatives into account. These measures were calculated on the test set and utilized to fairly compare model performance to ensure that the top-performing model was accurate, but also consistent in finding vitiligo for different cases.

CHAPTER 4

RESULTS AND DISCUSSIONS

This chapter presents the experimental results obtained from training and evaluating the proposed models on the vitiligo dataset. The performance of individual CNN models, ensemble learning techniques, and the custom MLP-based classifier were assessed using standard metrics—accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of each model's classification capability, especially in the context of medical image analysis where both false positives and false negatives can have significant implications.

Experimental results of training and testing the candidate models on the vitiligo dataset are descried in this chapter. The performance of a single CNN model, ensemble learning methods and an MLP based classifier developed by us was monitored using common metrics – accuracy, f1score, recall, precision. The metrics provide a rudimentary idea about the capability of each model to classify, specially in medical image processing where both the false positive and false negative can be significant

4.1 Individual CNN Model Performance

Five well-established convolutional neural networks were fine-tuned and tested on the vitiligo dataset. The objective was to evaluate their standalone performance and identify the most promising architectures for subsequent ensemble modeling.

- VGG16 achieved an accuracy of 81.33%, with high precision (89%) but relatively low recall (70%), indicating it often failed to identify all vitiligo cases, though its predictions were mostly correct when positive as shown in figure 4.1 and figure 4.2.
- ResNet50 performed significantly better, with 92.12% accuracy, 95% precision, and 74% recall, highlighting its improved balance between sensitivity and specificity 0as shown in figure 4.3 and figure 4.4.
- DenseNet121 also showed robust performance with 89.63% accuracy, 93% precision, and 66% recall, making it a strong candidate for feature extraction as shown in figure 4.5 and 4.6.
- InceptionV3 and EfficientNet underperformed, with 75.93% and 69.29% accuracy respectively as shown in figure 4.7 and figure 4.9. Their low recall and F1-scores—24% recall for InceptionV3 and 50% recall for EfficientNet as shown in figure 4.8 and figure 4.10—suggested that they were not reliable for detecting vitiligo cases and were excluded from further ensemble stages.

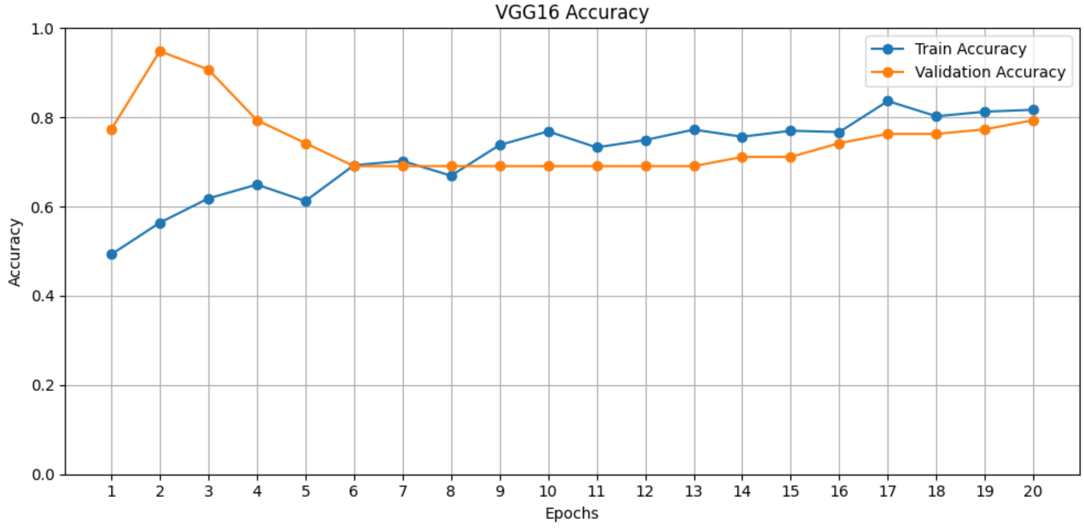


Figure 4.1 : VGG16 Accuracy Plot

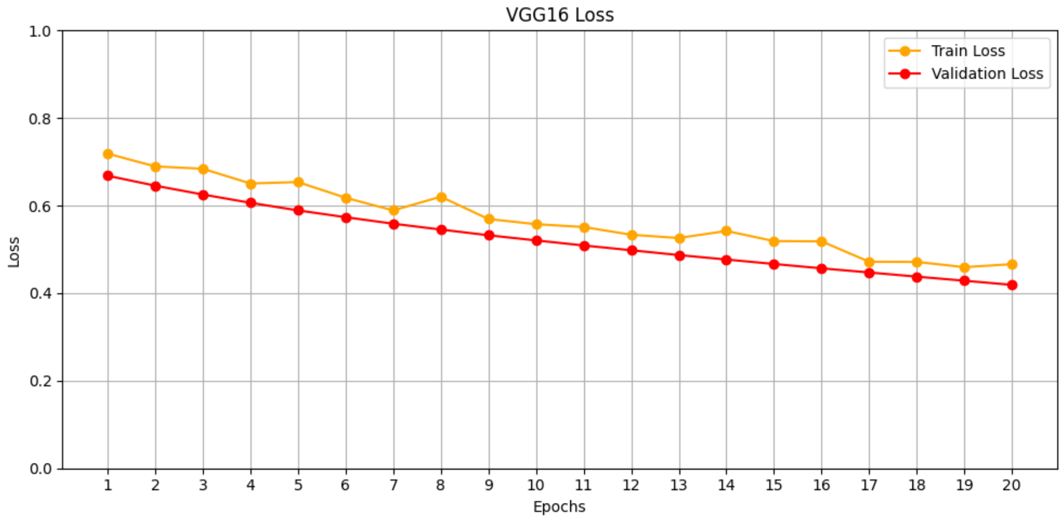


Figure 4.2 : VGG16 Loss Plot

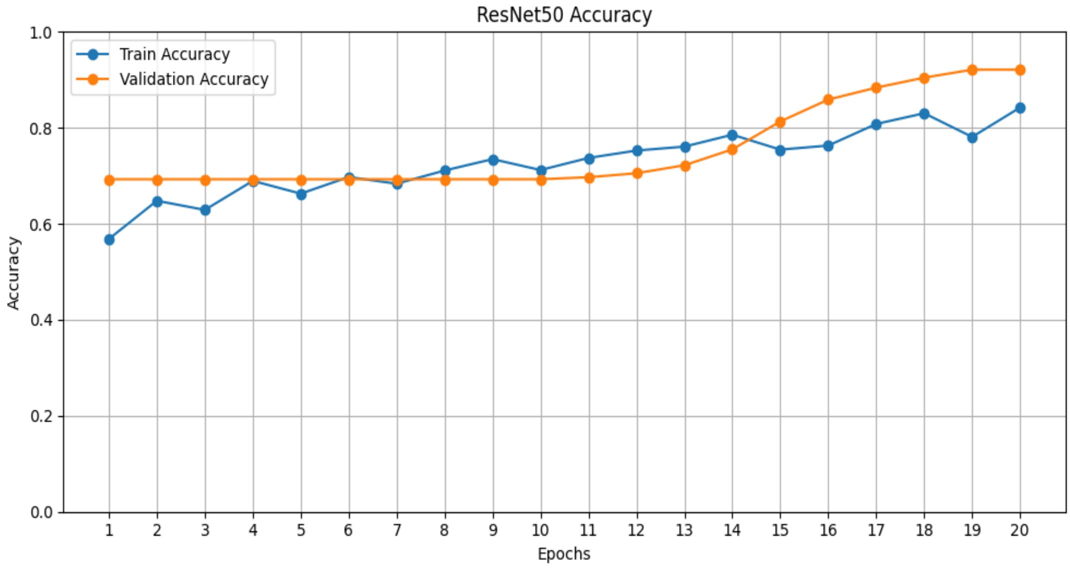


Figure 4.3 : ResNet50 Accuracy Plot

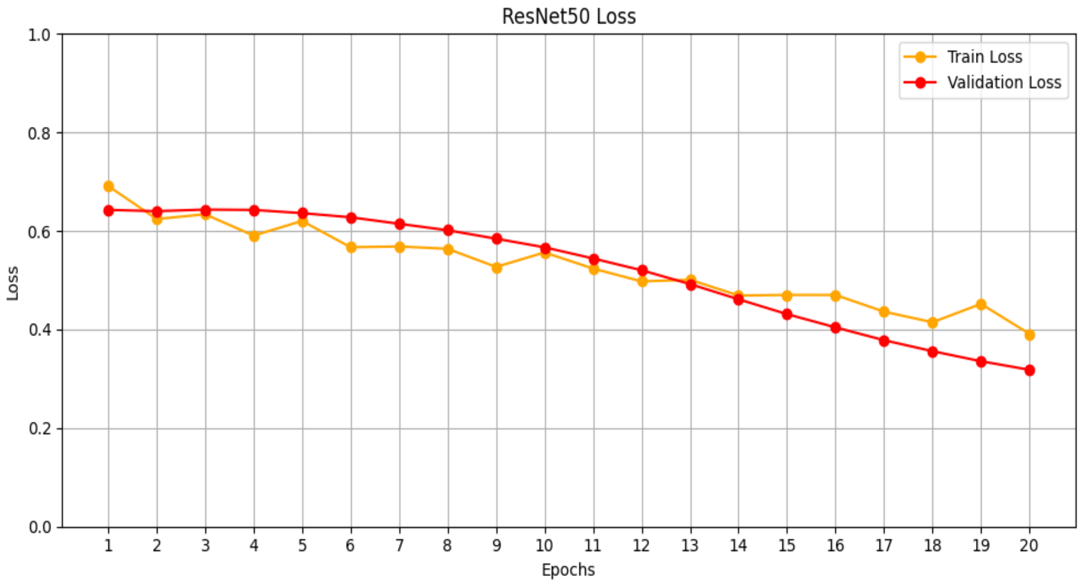


Figure 4.4 : ResNet50 Loss Plot

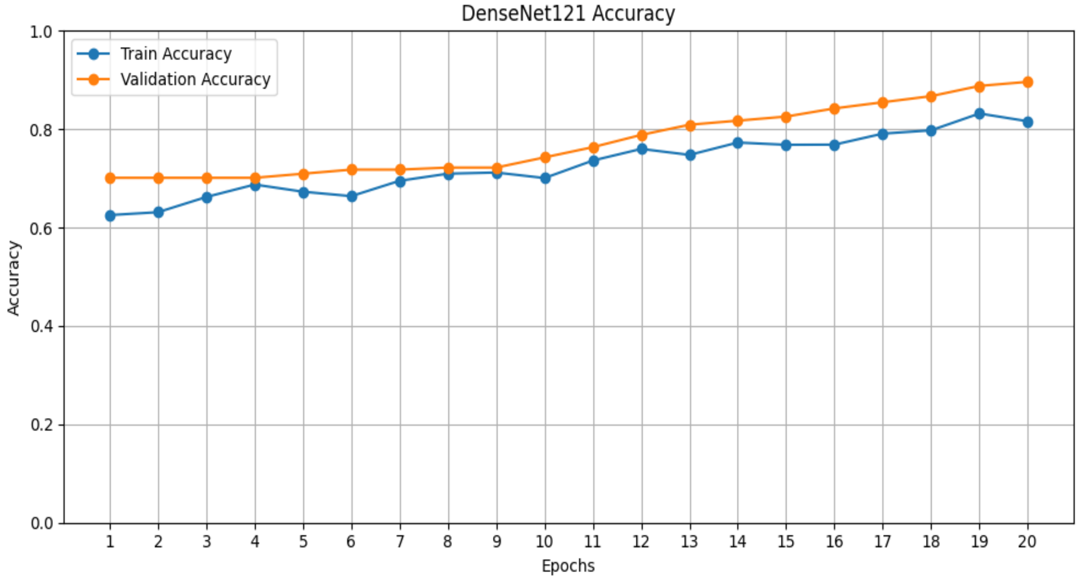


Figure 4.5 : DenseNet121 Accuracy Plot

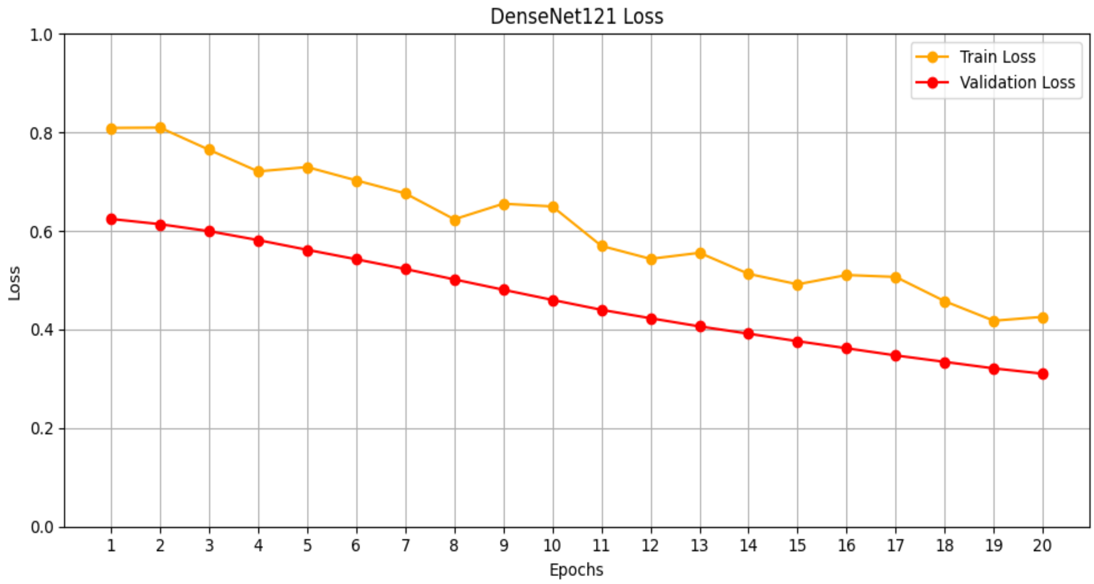


Figure 4.6 : DenseNet121 Loss Plot

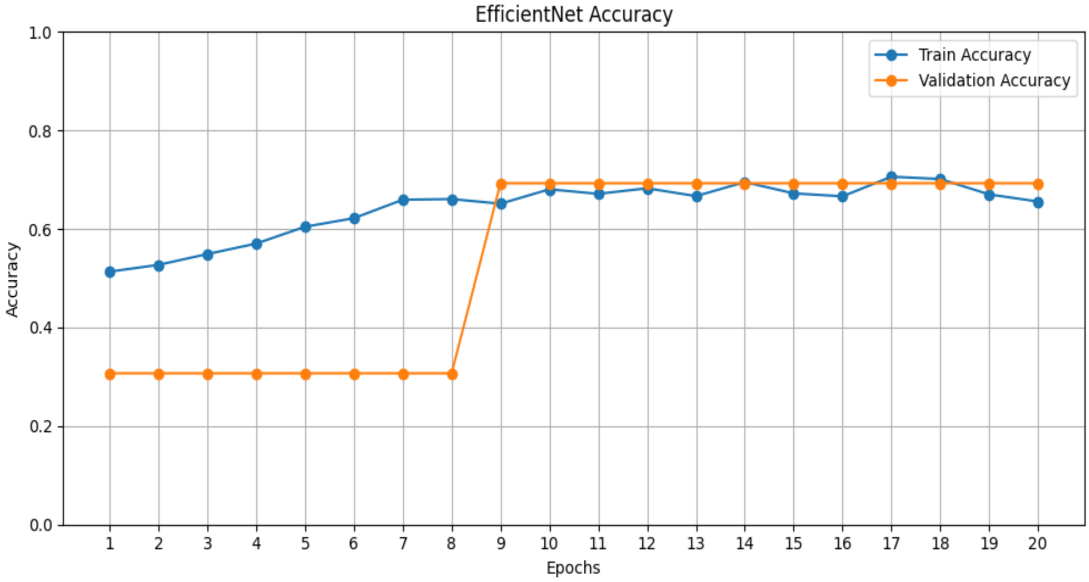


Figure 4.7 : EfficientNet Accuracy graph

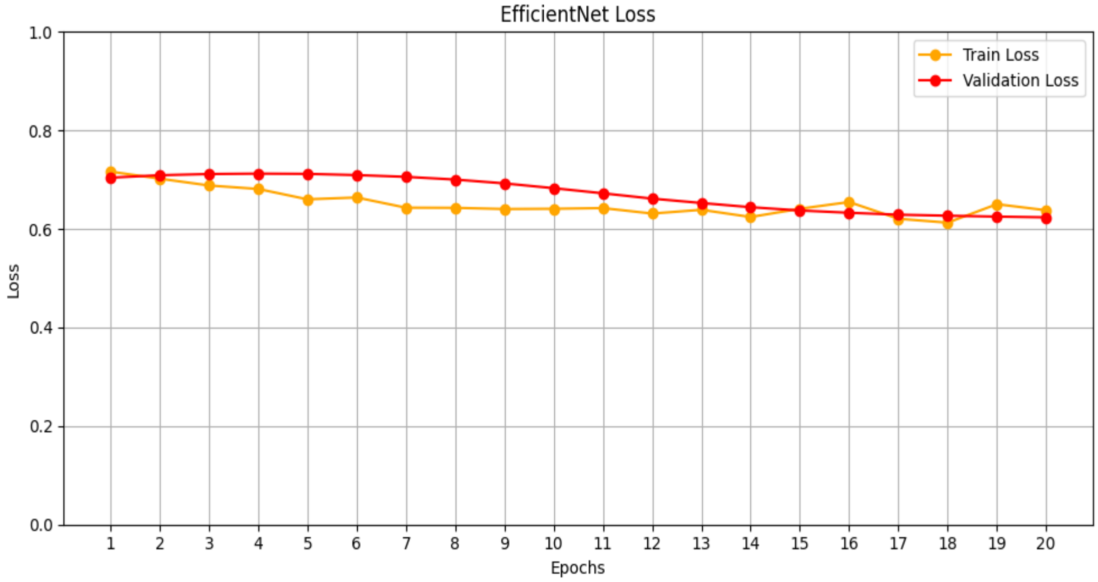


Figure 4.8 : EfficientNet Loss graph

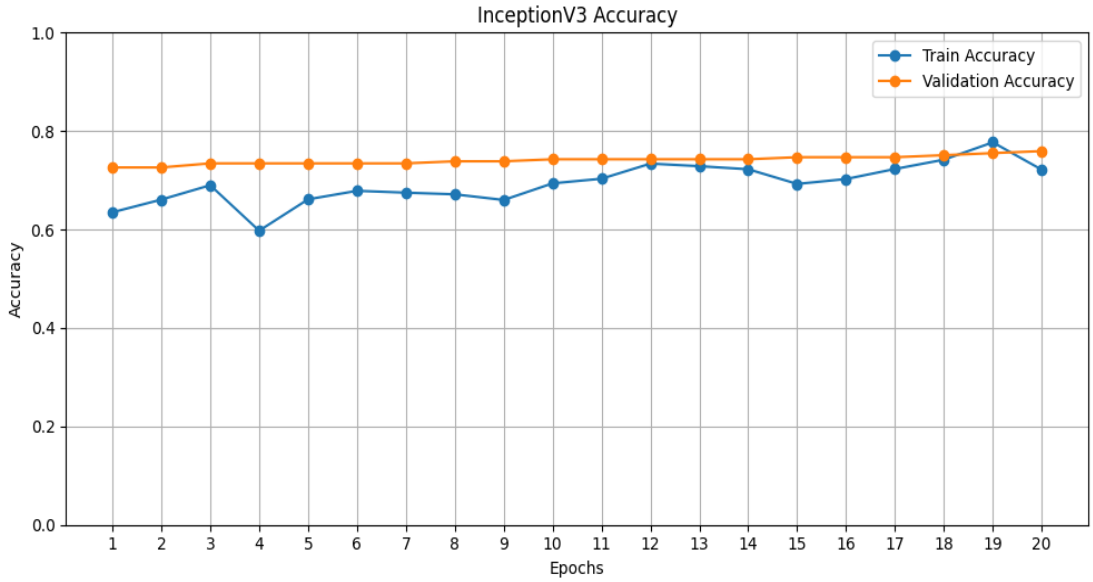


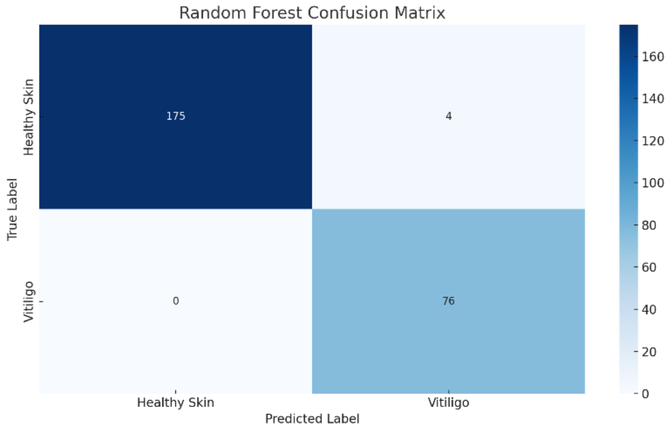
Figure 4.9 : InceptionV3 Accuracy graph

Figure 4.10 : InceptionV3 Loss graph

4.2 Ensemble Model Results

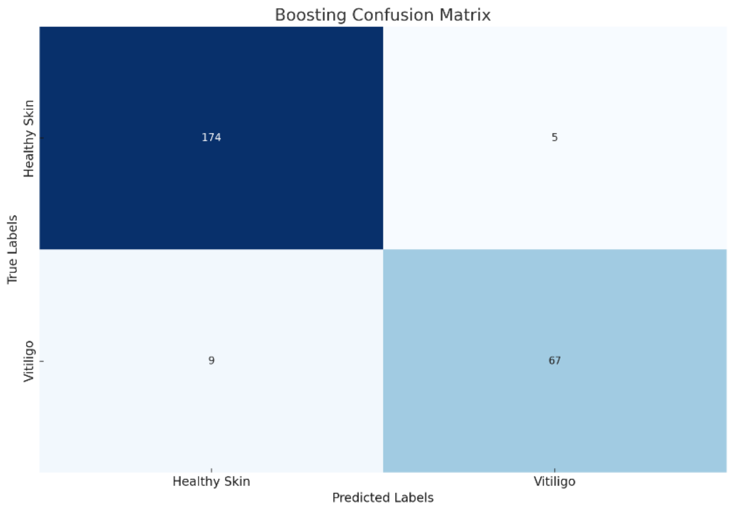
To overcome limitations of individual models, ensemble learning was explored using the top three performing CNNs: VGG16, ResNet50, and DenseNet121.

- Random Forest (bagging ensemble) achieved 98.43% accuracy, 97% precision, 98% recall, and 98% F1-score as shown in figure 4.11, indicating strong consistency and reliability.
- XGBoost (boosting) and Stacking both reached 94.51% accuracy, with similar scores for precision (94%) and recall (93%) as shown in figure 4.12 & figure 4.13. These models outperformed individual CNN but showed slightly lower generalization than Random Forest.



	precision	recall	f1-score	support
Healthy Skin	1.00	0.98	0.99	179
Vitiligo	0.95	1.00	0.97	76
accuracy			0.98	255
macro avg	0.97	0.99	0.98	255
weighted avg	0.99	0.98	0.98	255

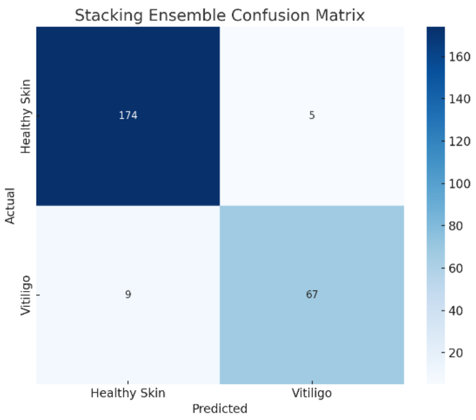
Figure 4.11 : Bagging (Random Forest) Classification Report



Boosting Classification Report:

	precision	recall	f1-score	support
Healthy Skin	0.95	0.97	0.96	179
Vitiligo	0.93	0.88	0.91	76
accuracy			0.95	255
macro avg	0.94	0.93	0.93	255
weighted avg	0.94	0.95	0.94	255

Figure 4.12 : Boosting (XGBoost) Classification Report



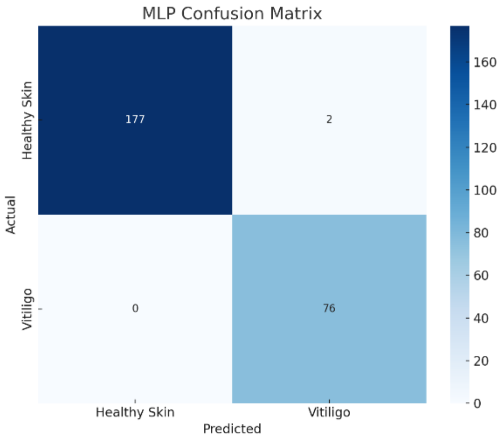
Stacking Ensemble Classification Report

	precision	recall	f1-score	support
Healthy Skin	0.95	0.97	0.96	179.0
Vitiligo	0.93	0.88	0.91	76.0
macro avg	0.94	0.93	0.93	255.0
weighted avg	0.94	0.95	0.94	255.0

Figure 4.13 : Stacking Classification Report

4.3 Custom MLP-Based Feature Fusion Model

The most significant performance improvement was observed in the custom MLP classifier, which took the concatenated deep features from VGG16, ResNet50, and DenseNet121 as input. The MLP model achieved 99.22% accuracy, 99% precision, 99% recall, and 99% F1-score as shown in figure 4.14, outperforming all other models. This demonstrated that feature-level fusion combined with a trainable neural classifier can capture complex relationships better than traditional ensemble voting schemes.



MLP Classification Report

	precision	recall	f1-score	support
Healthy Skin	1.0	0.99	0.99	179.0
Vitiligo	0.97	1.0	0.99	76.0
macro avg	0.99	0.99	0.99	255.0
weighted avg	0.99	0.99	0.99	255.0

Figure 4.14 : Proposed Model Classification Report

4.4 Discussion

The results validate that ensemble learning, particularly when implemented at the feature level rather than decision level, offers superior performance for vitiligo detection. Among the ensemble techniques, the MLP-based approach not only achieved the best accuracy but also maintained high recall, a critical factor in clinical diagnosis where false negatives are risky.

Furthermore, the disparity between precision and recall in models like InceptionV3 and EfficientNet reveals that high precision alone is not sufficient in medical applications. A balanced model that consistently identifies affected cases (high recall) is preferred. The findings also highlight the importance of selecting complementary models for ensemble construction. The feature maps from VGG16, ResNet50, and DenseNet121 provided diverse and rich representations, which proved advantageous when fused in the MLP classifier.

Overall, the experimental results show that carefully designed hybrid architectures combining multiple deep learning models can significantly enhance classification accuracy and reliability for vitiligo detection.

Table 2 : Perfomance Metrics

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
VGG16	81.33	89	70	81
ResNet50	92.12	95	74	85
InceptionV3	75.93	90	24	38
EfficientNet	69.29	35	50	41
DenseNet121	89.63	93	66	79
Bagging (Random Forest)	98.43	97	98	98
Boosting (XGBoost)	94.51	94	93	93
Stacking	94.51	94	93	93
Proposed Model	**99.22**	**99**	**99**	**99**

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This thesis offers a thorough deep learning-based approach to vitiligo classification using skin image data. Several pre-trained architectures of CNN VGG16, ResNet50, InceptionV3, EfficientNet, and DenseNet121—were assessed for the best model for this particular dermatology task. Out of these, ResNet50, VGG16, and DenseNet121 were the best models and were chosen for ensemble learning.

Ensemble strategies including bagging (Random Forest), boosting (XGBoost), and stacking were implemented using deep features extracted from the top CNNs. Additionally, a custom MLP-based feature fusion model was developed, which significantly outperformed all other methods.

The last results proved that the MLP fusion model performed best with 99.22% accuracy, 99% precision, 99% recall, and 99% F1-score, which showed a very reliable and generalizable diagnostic tool. In contrast, traditional models like InceptionV3 and EfficientNet underperformed, with lower recall and F1-scores, highlighting their limitations in vitiligo classification. Among the ensemble methods, Random Forest performed best with 98.43% accuracy, while XGBoost and stacking both reached 94.51%, albeit with slightly lower recall values.

These results clearly establish that combining multi-model features through a custom deep architecture provides superior performance over individual models and standard ensemble methods. The study validates the effectiveness of deep feature fusion and lays the foundation for future research in automated dermatological diagnostics.

5.2 Future Scope

While the proposed approach achieves excellent results, several areas remain open for further exploration:

Larger and more diverse datasets could be incorporated to improve the model's robustness across various skin tones, age groups, and lighting conditions. The current system performs binary classification (vitiligo vs. healthy). Future work can extend it to multi-class classification, covering other skin disorders like melasma, psoriasis, or eczema. Incorporating explainable AI (XAI) techniques such as Grad-CAM or SHAP can improve transparency and support clinical validation by highlighting decision-making regions in the images. Deployment in real-world clinical environments through mobile or web-based interfaces can help evaluate its practical utility and usability by dermatologists. Finally, integrating semi-supervised or self-supervised learning may help make the framework adaptable in settings where labeled data is limited.

BIBLIOGRAPHY

[1] Y. Zhang, Y. Cai, M. Shi, *et al.*, "The prevalence of vitiligo: A meta- analysis," *PLoS One*, vol. 11, e0163806, 2016.

[2] J. Z. Zhang, D. Luo, C. X. An, *et al.*, "Clinical and epidemiological characteristics of vitiligo at different ages: An analysis of 571 patients in Northwest China," *Int. J. Dermatol. Venereol.*, vol. 2, pp. 165–168, 2019.

[3] M. W. Linthorst Homan, P. I. Spuls, J. de Korte, *et al.*, "The burden of vitiligo: patient characteristics associated with quality of life," *J. Am. Acad. Dermatol.*, vol. 61, pp. 411–420, 2009.

[4] C. Salzes, S. Abadie, J. Seneschal, *et al.*, "The Vitiligo Impact Patient Scale (VIPs): Development and validation of a vitiligo burden assessment tool," *J. Invest. Dermatol.*, vol. 136, pp. 52–58, 2016.

[5] N. Elbuluk and K. Ezzedine, "Quality of life, burden of disease, co-morbidities, and systemic effects in vitiligo patients," *Dermatol. Clin.*, vol. 35, pp. 117–128, 2017.

[6] K. Ezzedine, V. Eleftheriadou, M. Whitton, *et al.*, "Vitiligo," *Lancet*, vol. 386, pp. 74–84, 2015.

[7] C. Y. Chen, W. M. Wang, C. H. Chung, *et al.*, "Increased risk of psychiatric disorders in adult patients with vitiligo: A nationwide, population-based cohort study in Taiwan," *J. Dermatol.*, vol. 47, pp. 470–475, 2020.

[8] E. Y. Gan, V. Eleftheriadou, S. Esmat, *et al.*, "Repigmentation in vitiligo: position paper of the Vitiligo Global Issues Consensus Conference," *Pigment Cell Melanoma Res.*, vol. 30, pp. 28–40, 2017.

[9] C. Bergqvist and K. Ezzedine, "Vitiligo: A focus on pathogenesis and its therapeutic implications," *J. Dermatol.*, vol. 48, pp. 252–270, 2021.

[10] Tanvir, S., Syed, S. A., Hussain, S., Zia, R., Rashid, M., & Zahid, H. (2024). Detection of Vitiligo Through Machine Learning and Computer-Aided Techniques: A Systematic Review. *BioMed Research International*, 2024, 3277546.

[11] L. Zhang, S. Mishra, T. Zhang, Y. Zhang, D. Zhang, Y. Lv, M. Lv, N. Guan, X. S. Hu, D. Z. Chen, and X. Han, “Design and Assessment of Convolutional Neural Network Based Methods for Vitiligo Diagnosis,” *Frontiers in Medicine*, vol. 8, Oct. 2021. doi: 10.3389/femd.2021.754202

[12] L. Guo, Y. Yang, H. Ding, H. Zheng, H. Yang, J. Xie, Y. Li, T. Lin, and Y. Ge, “A deep learning-based hybrid artificial intelligence model for the detection and severity assessment of vitiligo lesions,” *Annals of Translational Medicine*, vol. 10, no. 10, pp. 590–590, May 2022. doi: 10.21037/atm-22-1738

[13] B. Mondal, N. Das, K. C. Santosh, and M. Nasipuri, “Diagnosis of vitiligo using texture and colour features extracted from skin images,” in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Rochester, MN, USA, Jul. 2020, pp. 520–525. Doi: 10.1109/CBMS49503.2020.00104

[14] F. Zhong, K. He, M. Ji, *et al.*, "Optimizing vitiligo diagnosis with ResNet and Swin transformer deep learning models: a study on performance and interpretability," *Scientific Reports*, vol. 14, no. 1, p. 9127, 2024.

[15] V. Kantoria, S. Sharma, S. Bhushan, H. Saini, and R. Nijhawan, "Implication of convolutional neural network in the classification of vitiligo," *International Research Journal of Engineering and Technology (IRJET)*, vol. 7, no. 3, pp. 5170–5176, 2020.

[16] N. Bashar and M. Alsaid Suliman, *Vitiligo image classification using pre-trained convolutional neural network architectures, and its economic impact on health care*, M.S. thesis, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden, 2022.

[17] M. R. Thanka, E. B. Edwin, J. S. Duela, and V. Ebenezer, "Automated diagnosis of skin disease multi-class image classification using deep convolution neural network," *Journal of Green Engineering*, vol. 10, no. 10, pp. 7646–7660, 2020.