IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 May 2024, accepted 5 June 2024, date of publication 11 June 2024, date of current version 3 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3412791

RESEARCH ARTICLE

Genetic Algorithm Optimized Stacking Approach to Skin Disease Detection

ANANTHAKRISHNAN BALASUNDARAM[1,2], AYESHA SHAIK[1,2], B. ROHAN ALROY[2], AMOGH SINGH[2], AND S. J. SHIVAPRAKASH[2]

[1]Centre for Cyber Physical Systems, Vellore Institute of Technology (VIT), Chennai 600127, India
[2]School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Chennai 600127, India

Corresponding author: Ayesha Shaik (Ayesha.sk@vit.ac.in)

ABSTRACT Detection and treatment of skin diseases is a complicated process given the existence of about 3000 skin diseases. This adds complexity to the process of diagnosing skin diseases, highlighting the need for accurate detection to effectively treat the condition. Current deep learning-based skin detection tools generally focus on a narrow subset of skin diseases, work on relatively small datasets, and rarely achieve a Top-5 accuracy above 70%. Ideally, disease detection systems should possess the capability to detect and classify skin diseases, taking into account various environmental and situational factors. To overcome these challenges in detecting skin diseases, a deep learning-based system is proposed, utilizing an ensemble method with existing architectures to enhance performance through the integration of multiple models along with encoding to better adapt to varying inputs. The proposed deep learning-based system is trained on the DermNet dataset and uses a genetic algorithm optimized ensembling to enhance overall performance, resulting in a Top-5 accuracy of 74% on the DermNet dataset, a 5% improvement over the compared works. The system's performance is also evaluated using the HAM10000 dataset where the proposed system demonstrates an accuracy of 91.73%, a 2% improvement over the highest accuracy reported in the compared works.

INDEX TERMS Genetic algorithm, deep learning, ensembling, skin disease classification, dermnet.

I. INTRODUCTION

Despite the common occurrence of skin disease among people, many individuals do not consult a physician or an expert regarding their skin conditions. Skin diseases were responsible for more than 98,000 deaths worldwide and about 5 billion incidents in 2019 [1]. Skin diseases can also lead to disabilities, contributing to a decline in quality of life and life expectancy. The measure of years lost due to disability or low health is measured using a metric called Disability-Adjusted Life Years (DALY) which can be used to quantify the burden of a disease. In terms of DALY, skin diseases were responsible for 1.79% of the global burden of diseases in 2019 and are still a major cause of lifelong disability [1]. Table 1 given below indicates the increasing incidents and elevation in the number of DALYs and Deaths due to skin diseases globally over the years, as collected from the Global Burden of Disease Study [1].

TABLE 1. Number of incidents, DALYs and deaths due to skin diseases [1].

Year	Incidents (in Billions)	DALYs (in Millions)	Deaths (in thousands)
1990	3.127	32.918	49.281
1995	3.399	32.073	59.896
2000	3.660	33.331	61.621
2005	3.953	34.981	76.265
2010	4.261	38.668	95.835
2015	4.620	42.035	82.105
2019	4.859	42.883	98.521

Many approaches to skin disease detection have been implemented over the years; however, there are a few limitations that would have to be overcome to create a better system. The majority of existing solutions studied during the development of this system utilized a small dataset of images for training the deep learning models. A smaller

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeswari Sundararajan.

dataset is more prone to lower precision and accuracy due to underfitting and may not provide enough data to the model for a robust system. Another limitation of a few of the studied systems is that they were only built to detect a small number of skin diseases, decreasing their robustness. A few of the other systems are not capable of handling environmental and texture-based changes in the input image, which is a major factor as the environment and lighting conditions cannot always be controlled. Apart from this, a few systems were making use of models that may not be powerful enough to understand and detect from a comprehensive list of skin conditions. The vast number of skin diseases [2], coupled with the challenges posed by varying environments and small datasets, makes skin disease classification a critical issue.

The applications of deep learning have branched out into various domains, and one of them is healthcare. Deep learning in healthcare is an emerging domain owing to the better quality of healthcare that can be afforded by the implementation of deep learning-based approaches [3], [4], [5]. Some of the applications of deep learning include image analysis [6], [7], data analysis of healthcare data, and drug discovery. Deep learning systems built for detecting skin diseases could aid patients and doctors in increasing the chances of an early diagnosis, which could be followed up with a treatment to prevent mortality and decrease the chances of future complications and potential disability. Skin diseases can be detected based on their visual features, which suggests that a deep learning model could be used to detect skin diseases from images. This can potentially assist medical professionals in making better decisions. Such systems could also be used by patients to perform analysis at home and decrease the chances of a fatal outcome. A more accessible and robust system would go a long way in helping bridge the gap between medical facilities and professionals and would make testing and diagnosis more dependable, reliable, and accurate.

The primary objectives in developing the deep learning system were to use a comprehensive dataset that adequately covers the wide range of skin diseases present, a crucial factor given the large number of skin conditions affecting humans. The system must also possess the capability to detect multiple skin diseases, which improves its applications in medical diagnosis. The system would also need to demonstrate high accuracy to be dependable in real-world diagnosis.

The proposed system has applications for use in a medical environment, aiding medical professionals in determining skin diseases and guiding their diagnosis. An appreciable Top 5 accuracy demonstrates the effectiveness of the system to guide healthcare professionals by providing a list of highly probable diseases and assisting them in making the right diagnosis.

II. LITERATURE REVIEW

A method using a convolutional neural network (CNN) [8] is utilized to extract features from a dataset containing 100 images and perform skin disease detection through a support vector machine (SVM). The approach demonstrates the potential of using a CNN for feature extraction, which could allow for better results. Classification of skin diseases using architectures like MobileNetV2 [9], [10] uses images as input, and the output of the MobileNetV2 model is passed onto a Long Short-Term Memory unit, which then passes the data to a dense layer, which outputs the image classification, providing a unique approach which gives an accuracy of approximately 85%. Using an ensemble consisting of a CNN as a feature extractor and an artificial neural network (ANN) [11] to perform classification, which provides a better accuracy of 85.14%, A system to predict skin disease using ensemble data and feature selection [12] uses two approaches, one of which uses an ensemble approach with three methods along with Linear Discriminant Analysis (LDA), Passive Aggressive Classifier (PAC), Radius Neighbors Classifier (RNC), Bernoulli Naïve Bayesian (BNB), Naïve Bayes (NB), and Extra Tree Classifier (ETC) with gradient boosting, and the other approach uses a similar methodology but with a feature extractor to reduce the dataset. This system does not work on an image dataset and has a smaller number of classes. To detect skin cancer using a MobileNetV2 [13], transfer learning is utilized to classify a given image as cancerous or benign, achieving an accuracy of 98.2%.

Performing diagnosis of skin diseases using a custom model architecture that makes use of 2D Gabor filters and works on images of varying skin colors [14]. A custom convolutional model architecture to perform disease detection also uses principal component analysis with an accuracy of 90% [15]. The system had a high recognition rate for white blood cells but failed for a few image resolutions. A published work looks at a machine learning algorithm used for skin disease detection [16]. It discusses the usage and application of multiple machine learning models, which include Support Vector Machine (SVM), Logistic Regression, Random Forest, Naïve Bayes, and CNN, to identify the best model for the application. It concludes with the result that the CNN model achieved the highest accuracy of 96% compared to all the other algorithms. Implementing a ResNet [17] model to detect skin cancer [18] uses a transfer learning approach. A deep learning model is used with a ResNet model to predict images for skin cancer, but it has a lower accuracy of 57% in comparison to other disease-detecting models. Working on a limited labeled dataset, a novel transfer learning approach [19] is used to work on large datasets and achieve better accuracy. A Deep Convolutional Neural Network (DCNN) is used to classify images for skin cancer using double transfer learning. The approach achieved an F1 score of 89% for the base model and a 98% F1 score when used alongside a DCNN. This system, however, had a large training time due to the double transfer learning.

Using a convolutional neural network for skin disease classification [20], images are classified based on the

specific stages of melanoma, including lesion malignant, superficial spreading, and nodular melanoma. To tackle the lack of large datasets for the development of better models, DermGAN [21] is used, which proposes to use a generative adversarial network to synthesize images with a skin condition. A comparison study of machine learning and deep learning models differentiated the performances of the models [22]. The data was collected, preprocessed, features extracted, and then used for classification. The machine learning models compared were Bagged Tree Ensemble, K Nearest Neighbors, and Support Vector Machine, while the deep learning models used were VGG16, GoogleNet, and ResNet50. The most optimal performance was achieved by the Bagged Tree Ensemble model, which had an accuracy of 92.99%. The study helps us understand that good results can be reached using a testing dataset with less computational usage and better speed. The ML models, however, might not be able to perform equally well with larger datasets with higher dimensions. Performing skin disease detection using CNN requires a bigger dataset. To carry this out, a proposed method [23] increases the dataset by duplicating, which is then given as input to the model. This allows for better scalability and can handle higher dimensions. The average accuracy of the model across all the diseases is about 92%. Color and texture are important in classifying skin diseases, and the implementation of multiple machine learning algorithms [24] makes use of the color and texture features of the image to classify the skin disease, which is passed on to the decision tree and support vector machine. The features extracted from the images include entropy, variance, and the maximum histogram value of the Hue Saturation Value (HSV). As this approach depends on the color and texture features of the image, it may be prone to environmental factors.

The usage of multiple models within an ensemble method has been implemented in order to improve the accuracy of the individual models [25]. In an ensemble system, the combination of the models can improve the performance by aggregating the predictions of the models. There are various ensembling methods that can be employed, which can range from a voting system to a stacked system where the output of one model is given to another model [26]. The stacking of neural networks has been utilized to boost the accuracy in medical field applications. For instance, a stacking of convolutional neural networks is performed to improve the performance of the individual models for the classification of breast carcinomas [5]. Another method makes use of an ensemble system [27] consisting of an EfficientNet [28] and a custom neural network to classify images of the eye fundus for disease detection. To optimize an ensemble system, a method uses the genetic algorithm [29] to compute the ideal weights for multiple models and create a weighted voting ensemble. The ensemble optimized by the genetic algorithm was able to achieve higher performance compared to the other ensemble systems that were evaluated.

III. PROPOSED SYSTEM

The proposed system required the implementation of three modules for its functioning. The three modules are the preprocessing module, the prediction aggregator module for ensemble models, and the prediction module. The images are sourced from the DermNet dataset for the experiments. The preprocessing module performs appropriate preprocessing of the images from the dataset, which includes resizing, normalization, and image corrections. For models utilizing an ensemble approach, the prediction aggregator module collates all the results of the models in the ensemble to output the prediction. The prediction module provides the prediction from the proposed system for a given input image. The machine learning model for the proposed system consists of an ensemble model of ResNet50, DenseNet121, and a basic CNN. To identify optimal combinations of models, a genetic algorithm-based approach [30], [31] was used to effectively produce accurate ensembles. The genetic algorithm identified models that outperform individual performances when used together, which helps overcome the inefficiencies of single models.

Due to the requirement of a high-performance system, multiple models were identified and evaluated on the target dataset to choose the appropriate model. The steps performed for dataset preparation and model evaluation are described in Fig. 1. Firstly, the DermNet dataset is loaded, and then it performs preprocessing on the images, which includes encoder transformation and rescaling of pixels before being fed into the model. This ensures the uniformity of the input images to better fit the model. Data that is preprocessed is then split into training and test sets. The model is trained on the train dataset, while the test dataset is used to evaluate the model's performance after the training is completed. Training is then performed on the training data with one of the models. The loss and accuracy at every epoch in the training are saved. Once the models have been trained, they are then evaluated based on their accuracy on the test dataset. This metric is used in the next step of the genetic algorithm.

The genetic algorithm uses the performance metrics of the model to identify the best combinations of these models, which can enhance the overall performance of the system. Models selected by the genetic algorithm are then stacked together or put into an ensemble. The ensemble uses stacking, where every model's prediction is passed on to a dense neural network that makes a final prediction based on the predictions of each model.

For the selection of the models to be used, the genetic algorithm is implemented as described in Fig. 2. The fitness value for each ensemble is calculated by using test accuracy as a fitness metric for each ensemble. After calculating the fitness of the ensemble, the genetic algorithm is used to produce crossovers between any two ensembles. An ensemble is more likely to be included in a cross if it has a high fitness value. There are a few variables that are used to monitor and affect the functioning of the genetic algorithm. The variables are the random crossover rate,

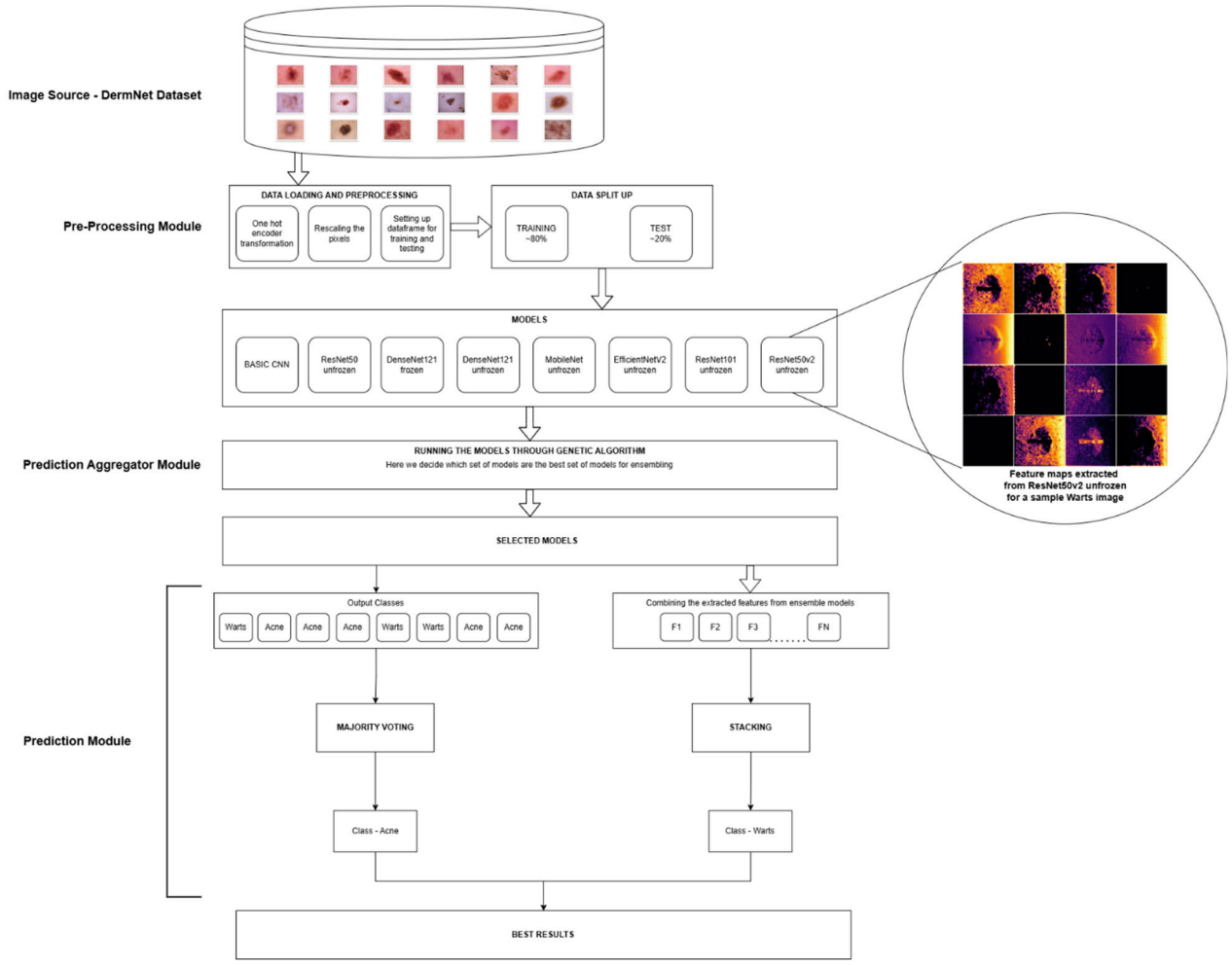

FIGURE 1. Block diagram of the proposed system.

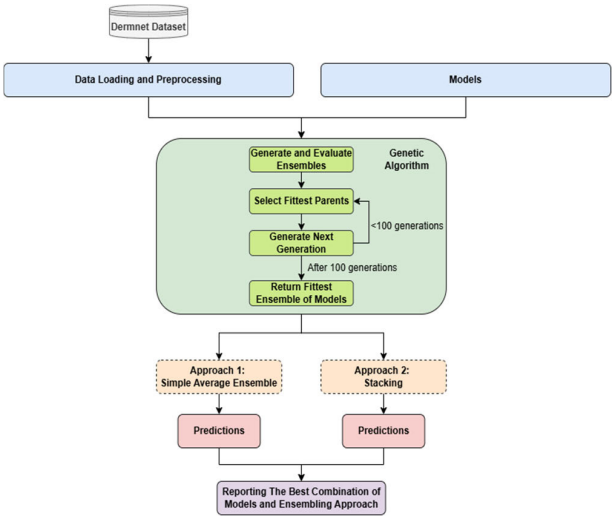

FIGURE 2. Flow diagram the proposed system.

which governs the probability that a random crossover takes place regardless of the fitness values, and the mutation rate, which randomly changes the values of the chromosomes (in this case, the model chosen as part of the ensemble). The stopping condition for the algorithm is set at the completion of 100 iterations of the genetic algorithm. At this point, the algorithm is stopped, and the best model is chosen. The algorithm therefore gives an ideal combination of the models in fixed iterations, which could maximize the performance of detecting skin diseases.

The final system architecture uses the models chosen by the genetic algorithm for ensembling.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental setup of the proposed work included the use of the TensorFlow library to help build a machine learning system and OpenCV to perform image augmentation and processing. The model is trained on the DermNet dataset obtained from the Dermatology Resource. The dataset consists of skin disease images in JPG format, and the images are split into 23 different classes based on the skin disease. There are over 19,000 images in the dataset, which were split into training and testing segments for the training and

TABLE 2. Initialization and hyperparameters.

Hyperparameter	Value
Batch Size	256
Optimizer	Adam
Loss function	Categorical Cross Entropy
Train Test split ratio	80% for train 20% for test

evaluation of the model. The classes present in the dataset are as follows:

- Acne and Rosacea
- Actinic Keratosis Basal Cell Carcinoma and other Malignant Lesions
- Atopic Dermatitis
- Bullous Disease
- Cellulitis Impetigo and other Bacterial Infections
- Eczema
- Exanthems and Drug Eruptions
- Hair Loss Alopecia and other Hair Diseases
- Herpes HPV and other STDs
- Light Diseases and Pigmentation Disorders
- Lupus and Connective Tissue Diseases
- Melanoma Skin Cancer Nevi and Moles
- Nail Fungus and other Nail Diseases
- Poison Ivy and other contact dermatitis
- Psoriasis Lichen Planus
- Scabies Lyme and other infestations and Bites
- Seborrheic Keratoses and other Benign tumors
- Systemic diseases
- Tinea Ringworm Candidiasis and other Fungal Infections
- Urticaria Hives
- Vascular Tumors
- Vasculitis
- Warts Molluscum and Viral Infections

Table 2 displays the parameters used for the training and testing of the models. The selection of batch size, optimizer, and loss function was followed by prior experiments, where the chosen values exhibited better performance for an equivalent number of epochs. Various values were explored in these experiments, including batch sizes of 16, 32, 64, 128, and 256. A batch size of 8 was initially considered but was omitted due to an increase in the training time which made the application of the genetic algorithm impractical. The options for the optimizer included the Adam optimizer, Stochastic Gradient Descent Optimizer, and the RMSProp optimizer. For the loss function, both categorical cross entropy and sparse categorical cross entropy were tested. The train-test split ratio was determined in accordance with standard practice, involving the random allocation of 80% for training and 20% for testing.

To assess the performance of each model and ensemble system trained, evaluation metrics including accuracy, precision, recall, F1 score, and Top 5 were used. Using multiple performance metrics offers a comprehensive understanding

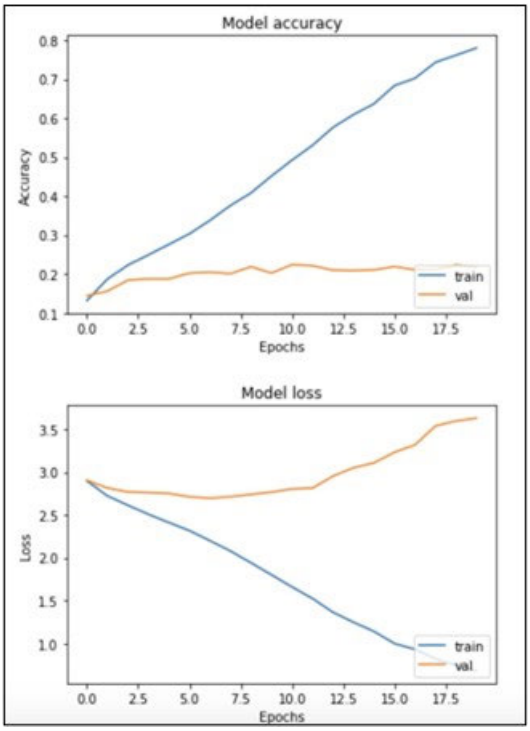

FIGURE 3. Accuracy and loss curve for basic CNN.

of the efficiency of each model and allows for an informed decision in the choice of models to be included.

A few reviewed works preferred to use a Top-5 statistic to help understand and grade the performance of the models trained and implemented. The Top-5 accuracy uses the top 5 predictions of the model based on their confidence scores and compares them to the actual value to calculate the statistic. For a more compatible comparison, the models trained during this work have been compared with other models based on their Top-5 accuracy along with the classical accuracy metric.

The models given below have been tested, and the model accuracy and model loss are recorded and plotted on a graph.

A. ALL LAYERS FROZEN

The following models had all their layers frozen, and only a few layers on the input and output sides were left unfrozen to allow modification of the layer weights.

Table 3 shows the results of evaluating the basic CNN model and Fig. 3 visualizes the plotting of the model accuracy and model loss for a basic CNN. Fig. 4 depicts the confusion matrix for the basic CNN model. The accuracy of the model increases in the training stage; however, the accuracy does not increase and stabilizes around 0.2 in the validation phase. This can be verified using the loss plots, where the loss decreases in the training phase but remains the same during the validation phase.

Table 4 shows the results of evaluating the ResNet50 model with the weights in all layers frozen and Fig. 5 visualizes the plotting of model accuracy and model loss for ResNet50 with weights in all layers frozen. Fig. 6 shows the confusion

FIGURE 4. Confusion matrix for basic CNN model.

TABLE 3. Metrics for basic CNN model.

Metric	Value
Accuracy	0.3448275923728943
Precision	0.3448275923728943
Recall	0.23938031494617462
F1 Score	0.3274094343656866
Top 5 Accuracy	0.6489255428314209

TABLE 4. Metrics for ResNet50 frozen model.

Metric	Value
Accuracy	0.17766116559505463
Precision	0.7864077687263489
Recall	0.020239880308508873
F1 Score	0.039464068684132526
Top 5 Accuracy	0.527486264705658

TABLE 5. Metrics for DenseNet121 frozen.

Metric	Value
Accuracy	0.3255872130393982
Precision	0.4426778256893158
Recall	0.26436781883239746
F1 Score	0.33103880104407796
Top 5 Accuracy	0.33103880104407796

matrix for Resnet50. The accuracy of the model in training and validation has a comparable increase, which is supported by the plot of the loss, which decreases with time in both phases, training and validation.

Table 5 shows the results of evaluating the DenseNet121 model with the weights in all layers frozen. Fig. 7 and Fig. 8 visualize the plotting of the model accuracy, model loss, and confusion matrix for a DenseNet121 with the weights of all layers frozen. The accuracy of the model in the training phase increases and reaches about 0.7, however, the accuracy in the validation phase remains stable around 0.2 throughout. The

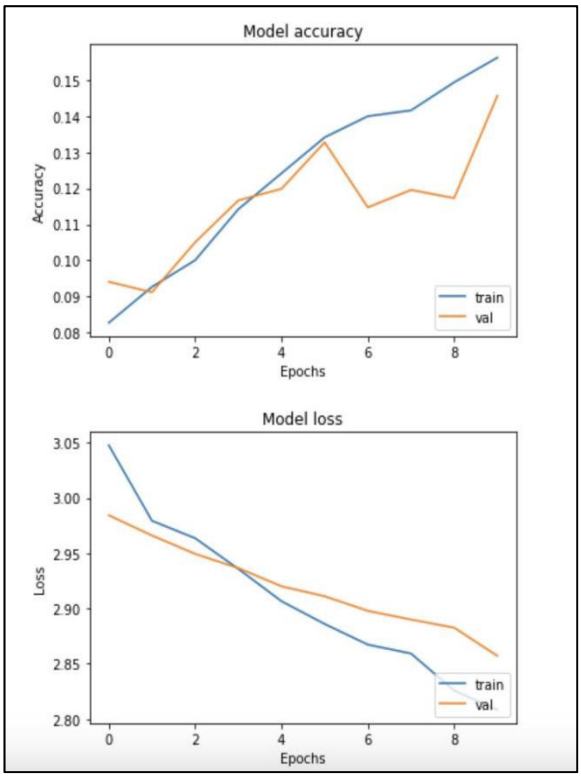

FIGURE 5. Accuracy and loss curve for ResNet50.

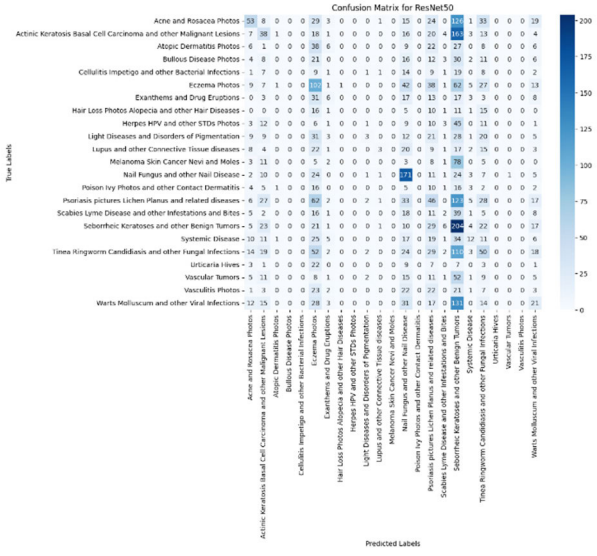

FIGURE 6. Confusion matrix for ResNet50.

loss also increases in the validation phase, whereas in the training phase, the loss decreases. This could be attributed to overfitting of the model.

B. FEW LAYERS UNFROZEN

For the models trained using a few layers unfrozen, the last 4 layers of the model were left unfrozen and trainable, allowing the modification of the weights to better adapt to the inputs.

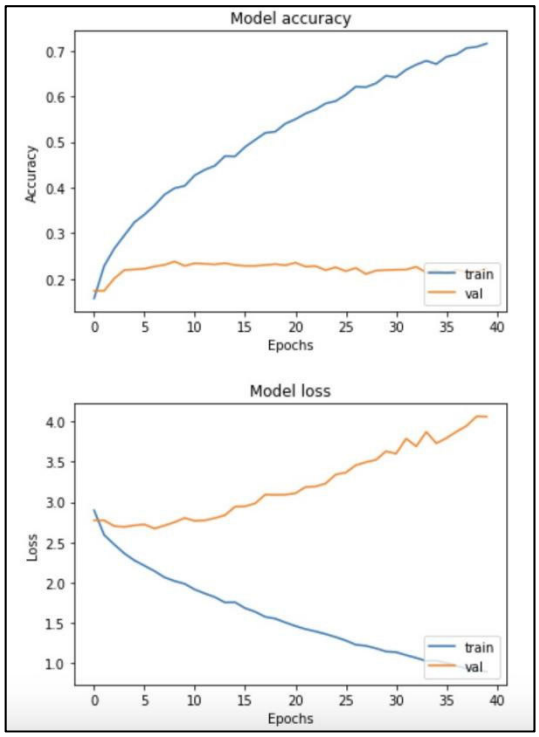

FIGURE 7. DenseNet121 frozen.

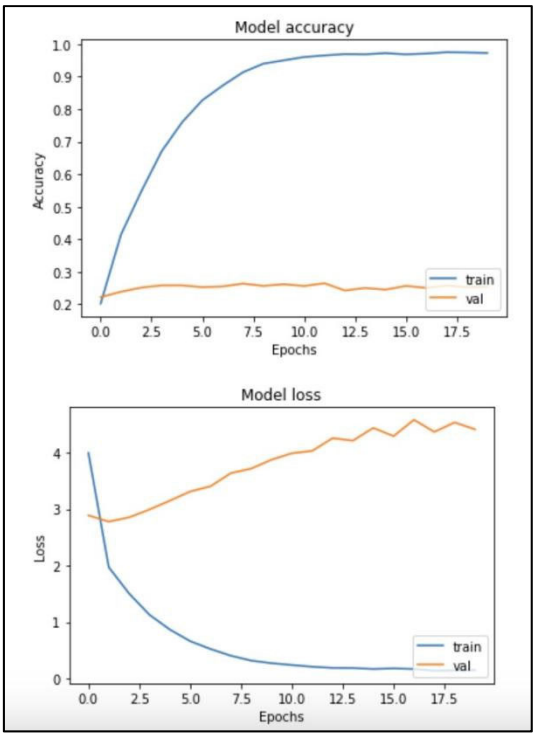

FIGURE 9. DenseNet121 unfrozen model.

FIGURE 8. Confusion matrix for DenseNet121 frozen model.

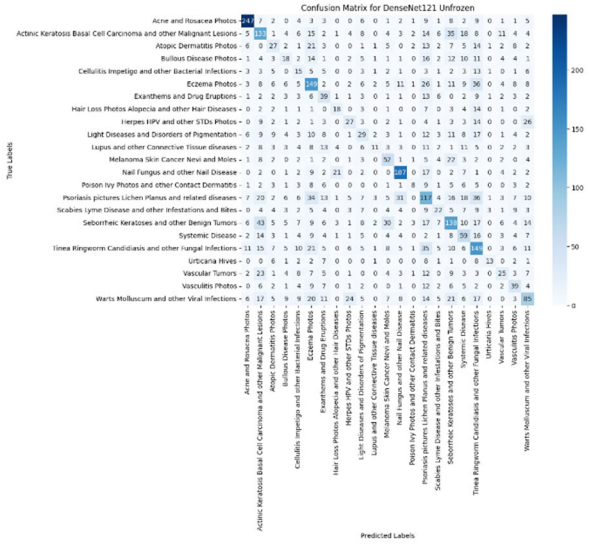

FIGURE 10. Confusion matrix for DenseNet121 unfrozen model.

TABLE 6. Metrics for DenseNet121 unfrozen model.

Metric	Value
Accuracy	0.40154922008514404
Precision	0.48531374335289
Recall	0.3633183538913727
F1 Score	0.4155473047234525
Top 5 Accuracy	0.7216391563415527

Table 6 shows the results of evaluating the DenseNet121 model with a few of the layers unfrozen. Fig. 9 and Fig. 10 visualize the plotting of the model accuracy, model loss, and confusion matrix for DenseNet121 with a few of the layers unfrozen. This implementation of the model does not improve the results compared to the previous models and still overfits. The accuracy of the training phase increases while the training loss decreases, but in the validation phase, the accuracy remains stable around 0.25 and the loss increases, indicating a possible overfitting.

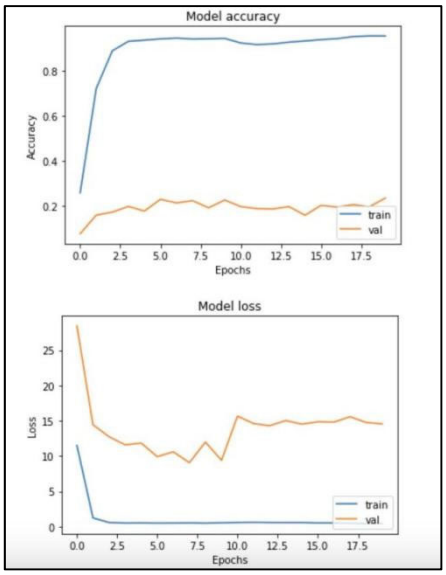

FIGURE 11. MobileNet unfrozen.

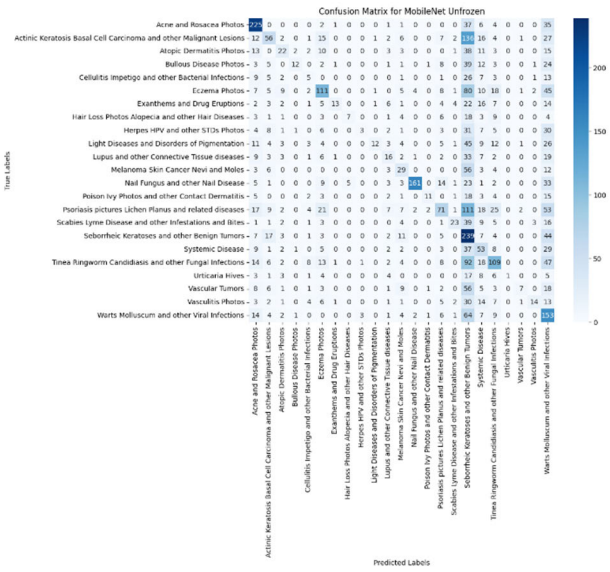

FIGURE 12. Confusion matrix for MobileNet unfrozen model.

Table 7 shows the results of evaluating the MobileNet model with a few layers unfrozen. Fig. 11 and Fig. 12 visualize the plotting of the model accuracy, model loss, and confusion matrix for the MobileNet model with a few layers unfrozen. During the training of the model, the accuracy of the model increases to about 0.9 and stabilizes, while the loss of the model decreases and remains stable. In the validation phase, the accuracy does not increase, but the loss of the model decreases. The model has a low validation accuracy.

Table 8 shows the results of evaluating the EfficientNetV2 model with a few layers unfrozen. Fig. 13 and Fig. 14 visualize the plotting of the model accuracy, model loss, and the confusion matrix of the EfficientNetV2 model with a few layers unfrozen. In this model training, the training accuracy of the model increases, but the validation accuracy does not increase. The model loss decreases in the training phase and validation phase, but due to the low accuracy of the model in the validation phase, it cannot provide very accurate results.

TABLE 7. Metrics for mobilenet unfrozen model.

Metric	Value
Accuracy	0.3380809724330902
Precision	0.34383806586265564
Recall	0.33533233404159546
F1 Score	0.3395319383009832
Top 5 Accuracy	0.6564217805862427

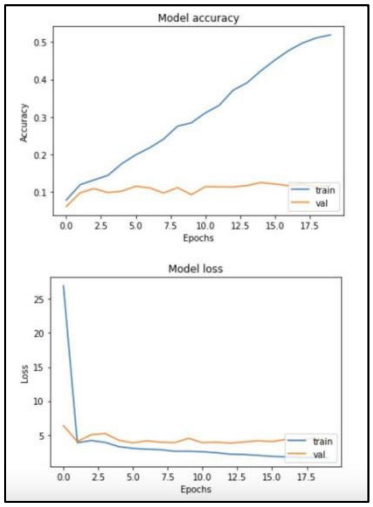

FIGURE 13. EfficientNetV2 unfrozen.

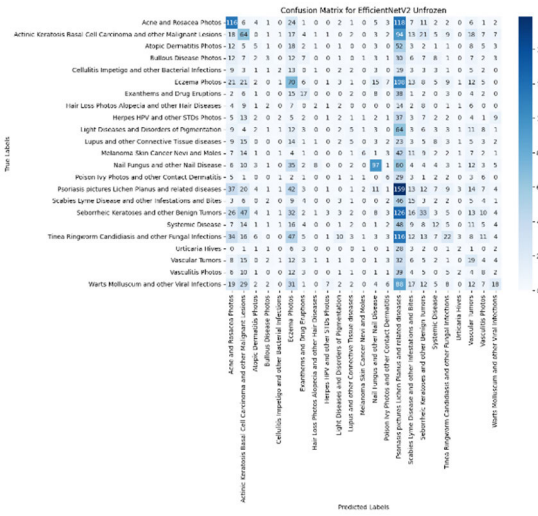

FIGURE 14. Confusion matrix for EfficientNetV2 unfrozen model.

TABLE 8. Metrics for EfficientNetV2 unfrozen model.

Metric	Value
Accuracy	0.17091454565525055
Precision	0.2517433762550354
Recall	0.09020489454269409
F1 Score	0.13281824560148067
Top 5 Accuracy	0.4670164883136749

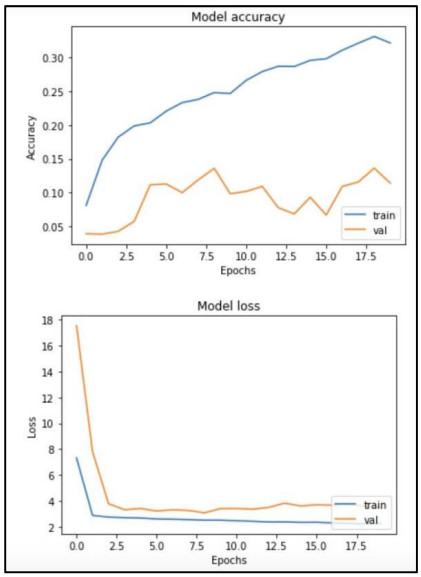

FIGURE 15. ResNet101 unfrozen.

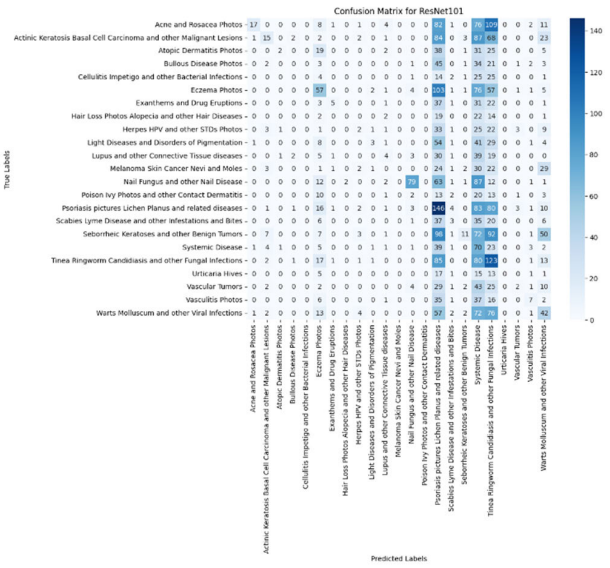

FIGURE 16. Confusion matrix for ResNet101 unfrozen.

Table 9 shows the results of evaluating the ResNet101 model with a few layers unfrozen. Fig. 15 and Fig. 16 visualize the plotting of the model accuracy, model loss, and confusion matrix of the ResNet101 model with a few layers unfrozen. The training accuracy of the model reaches approximately 0.35, with the validation accuracy being below the training accuracy. The train and validation loss decrease initially and remains stable. Due to the accuracy of the training and validation phases, the model may not be suitable for accurate results.

Table 10 shows the results of evaluating the ResNet50V2 model with a few layers unfrozen. Fig. 17 and Fig. 18 visualize the plotting of the model accuracy, model loss, and confusion matrix of the ResNet50V2 model with a few layers unfrozen. Examining the plots of the accuracy and loss of

TABLE 9. Metrics for ResNet101 unfrozen model.

Metric	Value
Accuracy	0.14692653715610504
Precision	0.2989247441291809
Recall	0.0347326323390007
F1 Score	0.0622341597525496
Top 5 Accuracy	0.4692653715610504

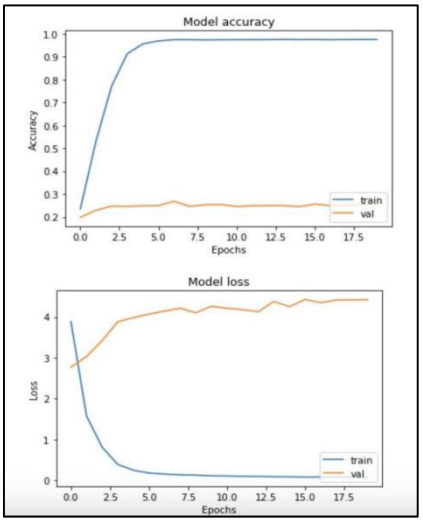

FIGURE 17. ResNet50V2 unfrozen.

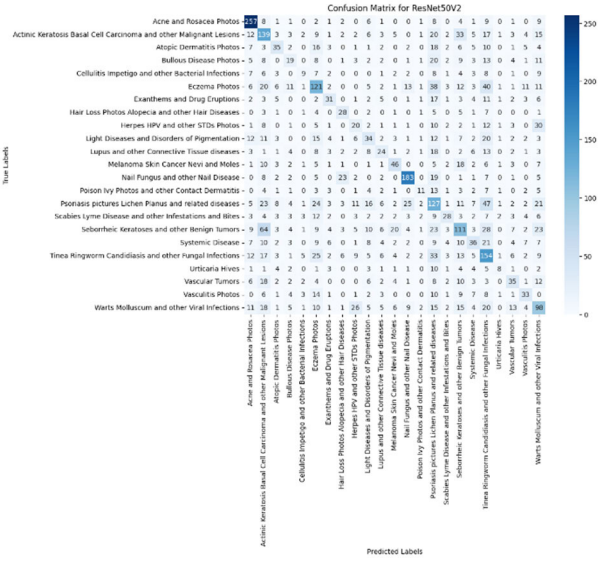

FIGURE 18. Confusion matrix for ResNet50V2 unfrozen model.

the model, it can be observed that the model is overfitting as the accuracy of the training phase is very high compared to the validation phase, and the loss also remains constant in the validation phase while in the training phase the loss decreases.

C. ENSEMBLE MODELS

To overcome the overfitting of the previously implemented models, the individual models were implemented in an

TABLE 10. Metrics for ResNet50V2 unfrozen model.

Metric	Value
Accuracy	0.3965517282485962
Precision	0.4872153974212646
Recall	0.34757620096206665
F1 Score	0.4057167742218094
Top 5 Accuracy	0.6989005208015442

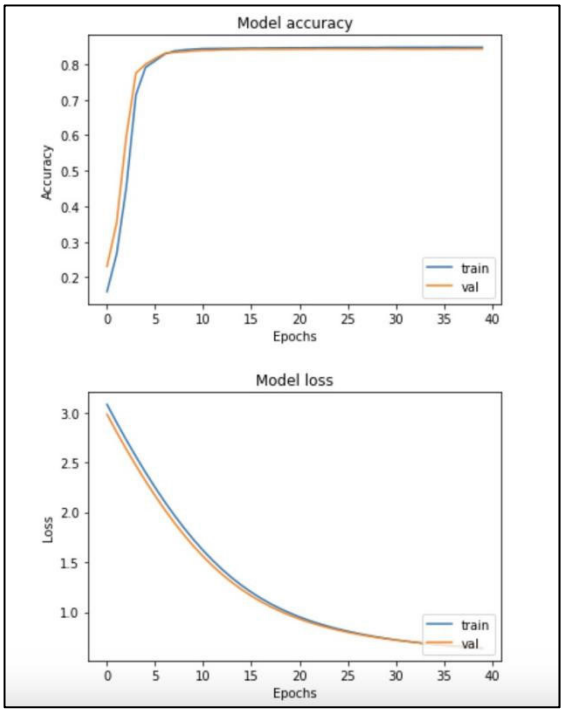

FIGURE 19. Stacking ResNetV2 unfrozen, DenseNet121 unfrozen and basic CNN.

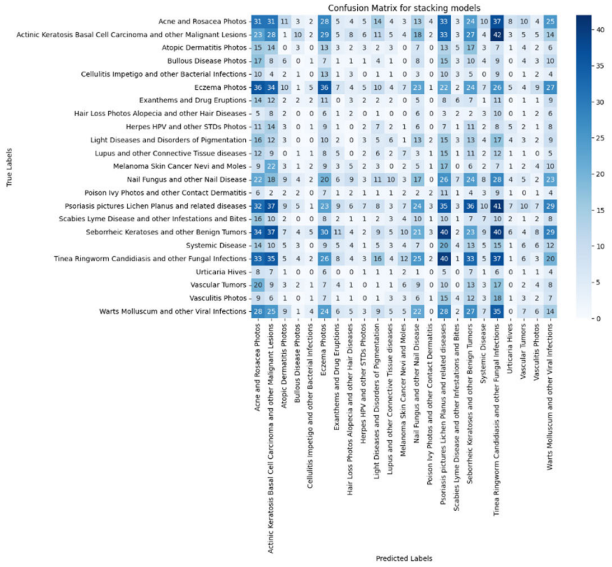

FIGURE 20. Confusion matrix for stacking models.

ensemble approach by combining the models, and one of the ensemble models was made according to the models chosen by the genetic algorithm.

TABLE 11. Metrics for stacking model.

Metric	Value
Accuracy	0.4537731111049652
Precision	0.746221661567688
Recall	0.29610195755958557
F1 Score	0.4239713860625446
Top 5 Accuracy	0.7353823184967041

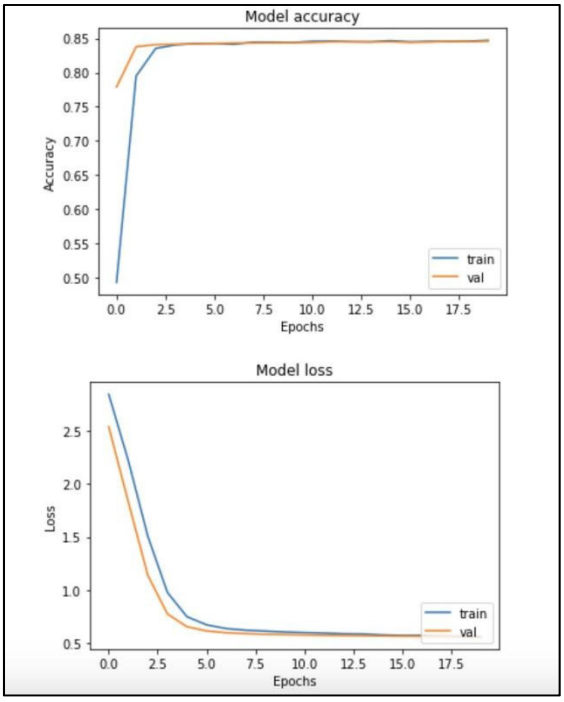

FIGURE 21. Stacking of genetic algorithm derived models - ResNet50V2 unfrozen, DenseNet121 unfrozen, basic CNN and ResNet50 frozen.

Table 11 shows the results of evaluating the stacking model. Fig. 19 and Fig. 20 visualize the plotting of the model accuracy, model loss, and confusion matrix of the stacking architecture consisting of ResNetV2 with a few layers unfrozen, DenseNet121 model with a few layers unfrozen, and a Basic CNN. This model shows better results than most of the models, as it does not demonstrate overfitting. The accuracy of the model in the training phase and validation phase is similar, showing a better fit. The accuracy of the model in the training and validation phases reaches around 0.8, whereas the loss decreases for training and validation to around 0.2.

Table 12 shows the results of evaluating the stacking of genetic algorithm derived models. Fig. 21 and Fig. 22 visualize the plotting of the model accuracy, the model loss, and confusion matrix of the stacking architecture, which consists of models selected by the Genetic algorithm. The models consist of ResNet50V2 with a few layers unfrozen, DenseNet121 with a few layers unfrozen, the Basic CNN, and ResNet50 with all layers frozen. In this ensemble method, the model has a good fit with the loss plots and the training plots for both phases, having a similar shape. The accuracy reaches

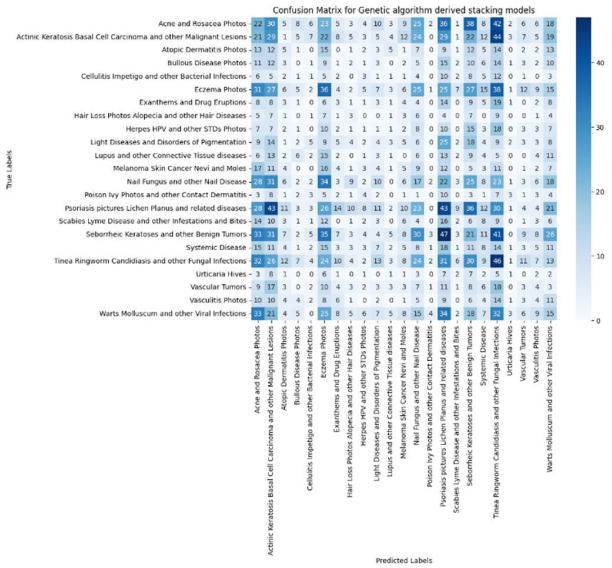

FIGURE 22. Confusion matrix for genetic algorithm derived models.

TABLE 12. Metrics for stacking of genetic algorithm derived models.

Metric	Value
Accuracy	0.4545227289199829
Precision	0.8125544786453247
Recall	0.23288355767726898
F1 Score	0.3620120393919364
Top 5 Accuracy	0.7338830828666687

about 0.85 for both phases, and the loss reaches about 0.5 for both phases.

The results from the training and testing of the model are recorded and provided in Table 13. The table contains details about the model and performance metrics, which are loss, accuracy, and the Top 5 accuracy of the model. Table 14 and Table 15 present a comparative analysis of the performance of the proposed models against those introduced by other researchers. The models were trained on datasets derived from DermNet and HAM10000, respectively.

Upon comparing the performance of the genetic algorithm optimized ensemble model with other existing works, it is evident that the optimized ensemble system yields better results in terms of accuracy and Top 5 accuracy for both the DermNet and HAM10000 datasets.

The compared systems consist of individual deep learning models as well as ensemble models, yet they demonstrate a lower accuracy compared to our proposed system. The decreased performance of individual deep learning models could be due to insufficient learning capacity, which can be solved by ensemble systems that benefit from the combination of multiple models. The suboptimal performance of ensemble systems can be due to inadequate optimization, resulting in lower accuracy despite the integration of multiple models.

Using an ensemble approach, we leverage the learning of multiple models with comparatively lower individual performance, combining them into an accurate ensemble model.

TABLE 13. Performance metrics of the implemented models and ensemble architectures trained ON DermNet dataset.

	Model	Loss	Accuracy	Top 5 accuracy
Base Case	CNN	2.8575	0.3448	0.6489
Using Pretrained Models as Feature Extractors	Resnet50	2.7414	0.1777	0.5275
	Densenet121	3.2352	0.3256	0.6387
Retraining Some Layers of a Pretrained Model	Densenet121 unfrozen	3.1213	0.4015	0.7216
	MobileNet unfrozen	11.1857	0.3381	0.6564
	Efficientnet v2 unfrozen	4.0056	0.1709	0.4670
	Resnet101 Unfrozen	3.3396	0.1469	0.4693
	Resnet50v2 unfrozen	3.1352	0.3966	0.6989
Simple Average Ensemble	Ensemble of Resnet50, Densenet121 unfrozen and CNN	2.1539	0.4538	0.7354
	Genetic Algorithm ensemble of CNN, Resnet50, Resnet50v2 unfrozen and Densenet121 unfrozen	2.0189	0.4545	0.7339
Stacking Based Ensemble	Stacking of Resnet50v2 unfrozen, Densenet121 unfrozen and CNN	1.9963	0.4578	0.7389
	Stacking of CNN, Resnet50, Resnet50v2 unfrozen and Densenet121 unfrozen	2.0202	0.4583	0.7401

TABLE 14. Comparison of models with best metric and the referenced models for DermNet dataset.

Model	Loss	Accuracy	Top 5 accuracy
Proposed Stacking of Resnet50v2 unfrozen, Densenet121 unfrozen and CNN	1.9963	0.4578	0.7389
Proposed Stacking of CNN, Resnet50, Resnet50v2 unfrozen and Densenet121 unfrozen	2.0202	0.4583	0.7401
AlexNet based CNN [33]	-	0.276	0.579
Ensemble of Pre trained VGG16, VGG19 and GoogleNet on OLE dataset [34]	-	0.311	0.69
Skin Image Search (Open access AI application) [35]	-	0.228	0.564

Given the large number of possible combinations of multiple models, we employ a genetic algorithm optimization, which

TABLE 15. Comparison of models with best metric and the referenced models for HAM10000 dataset.

Model	Loss	Accuracy	Top 5 accuracy
Proposed Stacking of CNN, Resnet50, Resnet50v2 unfrozen and Densenet121 unfrozen	0.8734	0.9173	0.9842
Ensemble learning model with Grey wolf optimization [36]	-	0.888	NA
Hand-crafted features with a 1D CNN [37]	-	0.8971	NA
EfficientNet [38]	-	0.8791	NA
Multiple features with Extreme Learning Machine Classifier [39]	-	0.8839	NA

provides us with the most suitable models for ensembling and is able to demonstrate better performance than the compared systems.

V. CONCLUSION AND FUTURE WORK

High accuracy skin disease detection systems are necessary due to the significance of the decisions being made. In the pursuit of an effective approach to skin disease detection, various models have been explored, along with ensembling and genetic algorithm techniques to determine optimized combinations of these models. Enhancing the effectiveness of the model in different environments is essential, and the image processing model is able to accomplish the task of improving the adaptability of the system. A stacking-based ensemble approach was implemented, merging multiple models to improve the overall performance of the models. The ensembled skin disease detection model demonstrated a viable approach for building high accuracy skin disease detection systems. a Top-5 accuracy of 74% was obtained over the DermNet dataset which wad 5% higher compared to other contemporary works. When evaluated over the HAM10000 dataset, the proposed system gave a good accuracy of 91.73% which was 2% higher than other works.

To enhance the achieved results, further exploration of the ensemble approach could involve implementing weighted-average-based ensemble techniques. Additionally, the ensemble approach could be refined by incorporating different neural network architectures. Furthermore, modifying individual parameters within model layers in future experiments may lead to better outcomes as these parameters were left unaltered through the course of the research work.

ACKNOWLEDGMENT

The authors would like to thank the Centre for Cyber Physical Systems and VIT Chennai management for their constant support and encouragement during this work.

REFERENCES

[1] *Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019 (GBD 2019) Results*, Institute for Health Metrics and Evaluation (IHME), Seattle, WA, USA, 2020.

[2] R. J. Hay, N. E. Johns, H. C. Williams, I. W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf, C. Michaud, C. J. L. Murray, and M. Naghavi, "The global burden of skin disease in 2010: An analysis of the prevalence and impact of skin conditions," *J. Investigative Dermatol.*, vol. 134, no. 6, pp. 1527–1534, Jun. 2014.

[3] Q. A. Al-Haija and A. Adebanjo, "Breast cancer diagnosis in histopathological images using ResNet-50 convolutional neural network," in *Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS)*, Sep. 2020, pp. 1–7.

[4] T. H. Arfan, M. Hayaty, and A. Hadinegoro, "Classification of brain tumours types based on MRI images using mobilenet," in *Proc. 2nd Int. Conf. Innov. Creative Inf. Technol. (ICITech)*, Salatiga, Indonesia, Sep. 2021, pp. 69–73.

[5] B. E. Bejnordi, G. Zuidhof, M. Balkenhol, M. Hermsen, P. Bult, B. van Ginneken, N. Karssemeijer, G. Litjens, and J. van der Laak, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *J. Med. Imag.*, vol. 4, no. 4, Dec. 2017, Art. no. 044504.

[6] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu, "Dense convolutional network and its application in medical image analysis," *BioMed Res. Int.*, vol. 2022, pp. 1–22, Apr. 2022.

[7] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 1, pp. 19–38, 2021.

[8] N. S. A. ALEnezi, "A method of skin disease detection using image processing and machine learning," *Proc. Comput. Sci.*, vol. 163, pp. 85–92, Jan. 2019.

[9] P. N. Srinivasu, J. G. Sivasai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM," *Sensors*, vol. 21, no. 8, p. 2852, Apr. 2021.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[11] J. Alam, "An efficient approach for skin disease detection using deep learning," in *Proc. IEEE Asia–Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, Dec. 2021, pp. 1–8.

[12] A. K. Verma, S. Pal, and S. Kumar, "Prediction of skin disease using ensemble data mining techniques and feature selection method—A comparative study," *Appl. Biochem. Biotechnol.*, vol. 190, no. 2, pp. 341–359, Feb. 2020.

[13] J. Rashid, M. Ishfaq, G. Ali, M. R. Saeed, M. Hussain, T. Alkhalifah, F. Alturise, and N. Samand, "Skin cancer disease detection using transfer learning technique," *Appl. Sci.*, vol. 12, no. 11, p. 5714, Jun. 2022.

[14] T. Shanthi, R. S. Sabeenian, and R. Anand, "Automatic diagnosis of skin diseases using convolution neural network," *Microprocessors Microsyst.*, vol. 76, Jul. 2020, Art. no. 103074.

[15] J. Huang, J. Li, Z. Li, Z. Zhu, C. Shen, G. Qi, and G. Yu, "Detection of diseases using machine learning image recognition technology in artificial intelligence," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–14, Apr. 2022.

[16] S. Bhadula, S. Sharma, P. Juyal, and C. Kulshrestha, "Machine learning algorithms based skin disease detection," *Int. J. Innov. Technol. Exploring Eng.*, vol. 9, no. 2, pp. 4044–4049, 2019.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[18] K. Das, C. J. Cockerell, A. Patil, P. Pietkiewicz, M. Giulini, S. Grabbe, and M. Goldust, "Machine learning and its application in skin cancer," *Int. J. Environ. Res. Public Health*, vol. 18, no. 24, p. 13409, Dec. 2021.

[19] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al-Shamma, M. A. Fadhel, J. Zhang, J. Santamaría, and Y. Duan, "Novel transfer learning approach for medical imaging with limited labeled data," *Cancers*, vol. 13, no. 7, p. 1590, Mar. 2021.

[20] V. R. Allugunti, "A machine learning model for skin disease classification using convolution neural network," *Int. J. Comput., Program. Database Manage.*, vol. 3, no. 1, pp. 141–147, Jan. 2022.

[21] A. Ghorbani, V. Natarajan, D. Coz, and Y. Liu, "DermGAN: Synthetic generation of clinical skin images with pathology," 2019, *arXiv:1911.08716*.

[22] S. Bandyopadhyay, A. Bhaumik, and S. Poddar, "Skin disease detection: Machine learning vs deep learning," Lincoln Univ. College, Kota Bharu, Kelantan, Malaysia, Tech. Rep., 2021, doi: 10.20944/preprints202109.0209.v1.

[23] M. Maniraju, R. Adithya, and G. Srilekha, "Recognition of type of skin disease using CNN," in *Proc. 1st Int. Conf. Artif. Intell. Trends Pattern Recognit. (ICAITPR)*, Hyderabad, India, Mar. 2022, pp. 1–4.
[24] K. V. Swamy and B. Divya, "Skin disease classification using machine learning algorithms," in *Proc. 2nd Int. Conf. Commun., Comput. Ind. 4.0 (C2I4)*, Bengaluru, India, Dec. 2021, pp. 1–5.
[25] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Mach. Learn.*, vol. 54, no. 3, pp. 255–273, Mar. 2004.
[26] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. MCS*, Berlin, Germany, 2000, pp. 1–15.
[27] J. Wang, L. Yang, Z. Huo, W. He, and J. Luo, "Multi-label classification of fundus images with EfficientNet," *IEEE Access*, vol. 8, pp. 212499–212508, 2020.
[28] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.
[29] A. Ekbal and S. Saha, "Weighted vote-based classifier ensemble for named entity recognition," *ACM Trans. Asian Lang. Inf. Process.*, vol. 10, no. 2, pp. 1–37, Jun. 2011.
[30] J. H. Holland, "Genetic algorithms," *Sci. Amer.*, vol. 267, no. 1, pp. 66–73, 1992.
[31] S. Mirjalili, "Genetic algorithm," in *Evolutionary Algorithms and Neural Networks*. Cham, Switzerland: Springer, 2018, pp. 43–55.
[32] H. Liao, Y. Li, and J. Luo, "Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 355–360.
[33] H. Liao, "A deep learning approach to universal skin disease classification," Dept. Comput. Sci., CSC, Univ. Rochester, Rochester, NY, USA, Tech. Rep. CSC-400, 2016.
[34] O. Zaar, A. Larson, S. Polesie, K. Saleh, M. Tarstedt, A. Olives, A. Suárez, M. Gillstedt, and N. Neittaanmäki, "Evaluation of the diagnostic accuracy of an online artificial intelligence application for skin disease diagnosis," *Acta Dermato Venereologica*, vol. 100, no. 16, 2020, Art. no. adv00260.
[35] L. Liu, X. Zhang, and Z. Xu, "An adaptive weight search method based on the grey wolf optimizer algorithm for skin lesion ensemble classification," *Int. J. Imag. Syst. Technol.*, vol. 34, no. 2, Mar. 2024, Art. no. e23049.
[36] A. Kumar, A. Vishwakarma, V. Bajaj, and S. Mishra, "Novel mixed domain hand-crafted features for skin disease recognition using multi-headed CNN," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.
[37] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using EfficientNets—A first step towards preventing skin cancer," *Neurosci. Informat.*, vol. 2, no. 4, Dec. 2022, Art. no. 100034.
[38] M. A. Khan, K. Muhammad, M. Sharif, T. Akram, and V. H. C. d. Albuquerque, "Multi-class skin lesion detection and classification via teledermatology," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 12, pp. 4267–4275, Dec. 2021.

ANANTHAKRISHNAN BALASUNDARAM received the master's degree in computer science and engineering from B. S. Abdur Rahman Crescent University, Chennai, India, and the Doctor of Philosophy (Ph.D.) degree in computer science and engineering from Anna University, India. He is currently an Associate Professor with the School of Computer Science and Engineering and is also associated with the Research Center for Cyber Physical Systems, Vellore Institute of Technology (VIT), Chennai Campus. He has an overall experience of 14 years of which he has over nine years of industrial experience working across MNCs like Cognizant Technology Solutions (CTS), Tata Consultancy Services (TCS), and iGATE Global Solutions, and five years of academic experience. His research interests include deep neural networks, computer vision, video analytics, image and video processing, artificial intelligence, data warehousing, data mining, healthcare intelligence, medical image analysis, and smart agriculture. He has received five best paper awards so far across international conferences. He has also received the Star Performer Award from Cognizant Technology Solutions and the Quality and Delivery Excellence Award from iGATE Global Solutions. He is also an active reviewer of reputed international SCIE journals of Elsevier, IEEE, and Springer. He has also served as the guest editor for special issues in a couple of SCI journals.

AYESHA SHAIK received the Doctor of Philosophy (Ph.D.) degree from IIITDM Kanchipuram. She is currently a Senior Assistant Professor with the School of Computer Science and Engineering and is also associated with the Research Center for Cyber Physical Systems, Vellore Institute of Technology (VIT), Chennai Campus. Her research interests include image processing, digital image processing, watermarking, and deep learning.

B. ROHAN ALROY is currently pursuing the Bachelor of Technology degree in computer science and engineering with a specialization in artificial intelligence and machine learning with the School of Computer Science and Engineering, Vellore Institute of Technology, Chennai. His research interests include technology and AI/ML in the fields of medicine, finance, and computer vision.

AMOGH SINGH is currently pursuing the Bachelor of Technology degree in computer science and engineering with a specialization in AI and robotics with Vellore Institute of Technology, Chennai. His research interests include AI/ML, drones, and robotics.

S. J. SHIVAPRAKASH is currently pursuing the degree in computer science with a specialization in artificial intelligence and machine learning with Vellore Institute of Technology, Chennai. He is exploring various research works on various topics connected to deep learning.