CAPSTONE PROJECT REPORT

Skin Lesions Classification with advanced Computer Vision Models

Submitted in partial fulfillment of the requirements
for the award of the degree of

Master of Technology
(2024-26, 3rd Sem)

Under the Supervision of:
Dr. Rajesh Kumar Shrivastava
Associate Professor (SCSET)

Submitted by:
Akhil Sibi (S24MTCG0001)

Plot Nos 8-11, TechZone 2, Greater Noida, Uttar Pradesh 201310

2024-26

ABSTRACT

A skin lesion is basically an abnormal mark on the skin, and its appearance may uniquely differ in size and shape compared to the surrounding area of the normal skin. Skin lesions can come in various appearances and can be associated with various skin cancer conditions such as the following: actinic keratoses (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vas). Most dermatologists view these skin lesions under good lighting conditions, which sometimes can result in human error in identifying the exact skin condition from the lesions. Thus our research interest is on the development of the classification of skin lesions using advanced computer vision models such as GhostNetV2_100, EfficientNetV2_b0_r224, FastVit_t8.apple, ShuffleNet_v2_x1_0, LeViT_conv_128s.fb_dist, MobileVit_xxs, MobileNetV4_conv_small.e2400_r224, ghostnetv3_100 , mobileone_s0.apple_in1k , resnext50_32x4d.a1h_in1k , resnetv2_50.a1h_in1k, and efficientformerv2_s0 . These models are chosen as they can capture the delicacies of dermatological images and address the issues of color, texture, and differences in shapes of skin lesions in an efficient and effective manner for image identification tasks. We used complex data preprocessing techniques to improve the quality of images and further expand the HAM10000 dataset that contains more than 10,000 pictures of various skin lesions in the dermatoscope. Hyperparameter optimisation, more strict training algorithms of the models deliver accuracy. Compared on the basis of testing accuracy, GhostNetV2_100 is the promising model in this area of real-time categorization of skin lesions with the test accuracy rate of 98.60%. We also explainable AI techniques such as GradCAM etc. so that the model internal predictions on the sample image can be visualized in order to understand which parts of the image the model focuses on. This research will also be valuable for building dependable and accessible methods of automation for the diagnosis of skin lesions that might be useful for doctors in the early detection and improve the treatment of the patient.

Keywords : Vision transformer , Convolutional Neural Networks, Ghostnet, FastViT, EfficientFormer, MobileNet, ResNet

TABLE OF CONTENTS

INTRODUCTION

Melanoma poses a great burden on health care systems around the world since it is the deadliest form of skin cancer. It calls for precise classification of such skin cancer conditions to ensure proper treatment is given, and best outcomes are optimized for patients. The conventional existing methods of diagnosis have relied heavily on subjectivity and interobserver variability since most usage depended on the visual inspections by the dermatologists. This can lead to delays in diagnosis, particularly where features are ambiguous or uncommon, thereby prejudicing the outcome of treatments.

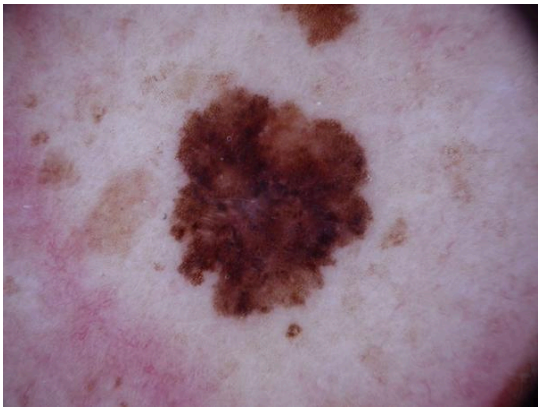


Figure 1: Melanoma image from ham10000 dataset

So, Deep learning can be stated as that sub-field of artificial intelligence that has been dominant in the area of medical image classification. With the help of deep learning models inspired by the structure and function of the human brain, promising solutions for automated skin lesion classification are surfacing such as CNNs, which can learn very complicated patterns and features from humongous datasets, therefore making them truly accurate in the classification of different types of skin lesions.

Computer Vision field saw a tremendous leap lately with the appearance of Vision Transformers[1](ViT) in 2021. Unlike CNNs that process images in succession, ViTs base their approach on an attention mechanism that captures long-range dependencies and a global context within an image. Therefore ViT model architecture captures the global context in the images due to their self attention mechanism and clearly ViT is better at capturing the information in the images than the cnn architectures. This allows them to pinpoint salient elements and the interdependencies thereof all across the image, thereby improving their performance on a multitude of vision tasks, including skin lesion classification. As such, CNNs and other ViTs will remain

dominant in medical image classification as they offer valuable inputs that could translate to more informative and better decisions for doctors.

Our research study explores the popular advanced computer vision models for the classification of skin lesions in the dataset. We pay attention to the most comprehensive as possible variability of models' architectures: GhostNetV2_100, EfficientNetV2_b0_r224, FastVit_t8.apple, ShuffleNet_v2_x1_0, LeViT_conv_128s.fb_dist, MobileVit_xxs ,MobileNetV4_conv_small.e2400_r224, ghostnetv3_100 , mobileone_s0.apple_in1k , resnext50_32x4d.a1h_in1k , resnetv2_50.a1h_in1k, and efficientformerv2_s0. Therefore this comparative study was done by measuring their efficiency, benchmark image recognition success and, secondly, their ability to represent the fine details of dermatological images.

We use the HAM10000 dataset, one of the most immense collections of dermoscopic images that vary from many other kinds of skin diseases. Such images are available: Actinic Keratoses (akiec), Basal Cell Carcinoma (bcc), Benign Keratosis-like Lesions (bkl), Dermatofibroma (df), Melanoma (mel), Melanocytic Nevi (nv), and Vascular Lesions (vasc), which becomes a good basis for training and testing these advanced computer vision models. Applying high-end data preprocessing like image quality enhancement and dataset augmentation techniques would result in deeply robust and generalizable models. In our research, rigorous experimentation and comparative studies will most probably generate the most effective and efficient deep learning architectures to classify skin lesions both accurately and in an efficient manner.

This study contributes to the long list of research in computer-aided diagnosis toward the realization of accurate and accessible tools for the automated scrutiny of skin lesions. We envision a future in which tools such as those developed here would be created and made available to doctors to better bring about more accuracy in the process of diagnosis and speed the treatment decision-making process, thus contributing to better patient outcomes in this war against skin cancer.

PROBLEM STATEMENT

The main obstacle in dermatology continues to be the accurate and timely identification of a skin lesion, seriously impaired patient care and treatment of the disease, due to many factors:

1. *Visual Examination*: The visual examination conducted by conventional diagnosis has very important inter-observer variability. Lesion morphology may be complex and subtle, and in some cases, it may be quite tough to diagnose, hence leading to diagnostic inconsistencies and huge therapeutic delays specially in equivocal cases. With the rising incidence of skin cancers, and heterogeneity in the disease of the skin lesion, it creates a heavy burden on the health care systems. The right classification of many diseases, including melanoma, basal cell carcinoma, and benign lesions, is very important for the proper treatment.

2. *Lack of Expert:*. Not all locations or cities can offer ready access to expert dermatologists, especially in remote or less developed locations. Access to experts would be missed; this may lead to over-delayed diagnosis, and disastrous consequences towards the patients in such places.

3. *Need for Real-Time Diagnostic support*: The world today is in strong need of such real-time diagnostic support technologies through which doctors can make proper and efficient decisions while assessment. Such techniques boost the confidence of diagnostic accuracy, provide early interventions, and improve patient outcomes.

4.*Dermatological Image Complexity*: Images in dermatology are fairly complex as they contain considerable differences in color and texture, form, and size. Capturing this fine information and then interpreting them is a daunting task for an automated diagnostic system.

This research work addresses the above issues by training and testing advanced computer vision models for accurate classification of skin lesions which is crucial for accurate diagnosis of skin cancer condition. We'll develop robust and accurate models capable of:

- The objective diagnosis and consistency in the process above eliminated subjectivity.
- Correct classification of a very large number of skin lesions.
- Providing more user-friendly and accessible tools to facilitate access to diagnostic support.
- Real-time analysis allows for timely interventions.
- Captures the dermatological images with a subtle intensity.

By completing these objectives, this comparative study contributes to the general development of computer-assisted diagnosis in dermatology with the hope that this would improve patient care and treatment results in the battle against skin cancer.

LITERATURE REVIEW

Methodology

HAM10000 [3] Dataset is a very popular dataset in the field of medical imaging and it is very good in studying different types of skin lesions that contribute to various skin conditions. Classification of skin cancer lesions has seen meteoric research study of techniques such as machine learning and deep learning, especially Convolutional Neural Networks. Very recently, it has been proposed by Amirreza Mahbod et al. [4] to propose a hybrid model architecture combining the pre-trained model like AlexNet, VGG16 and ResNet-18 along with SVM classifiers, that would their performance on the ISIC 2017 dataset. Similarly, Romero Lopez et al. [6] applied a transfer learning approach on the VGGNet architecture towards success in sensitivity on the ISIC Archive dataset. Bhuvaneshwari Shetty et al. [7] carried out data augmentation of images in the HAM10000 dataset with CNNs. The results in that experiment depicted that the accuracy rate was above the traditional machine learning method. Zhang et al. [5] designed an ARL-CNN model specifically for better feature discrimination due to the incorporation of self-attention mechanisms in skin cancer lesions detection.

Research Discovery

Improved deep learning architectures opened the door for promising results in skin cancer lesions classification. The hybrid model by Mahbod et al [4], classified melanoma with an AUC of 83.83% and therefore this model is efficient for classifying dermoscopic images. Lopez et al. [6] have reported increased sensitivity using VGGNet with transfer learning. It shows that fine-tuning CNNs with the dermoscopic images yields better performance than training from scratch . Shetty et al. [7] have shown that data augmentation techniques considerably improve model robustness to most lesion types, while accuracy in image classification is at 95.18% . Zhang et al.'s [5] ARL model presented results using an attention-based approach for focusing on lesion-specific locations, targeting for high accuracy without a reliance on big datasets.

Research Gaps and Limitations

While progress has been made, several areas persist. Mahbod et al. [4] indicated that limited datasets and visual artifacts reduce the generalization of CNNs. The HAM10000 dataset [3], despite

its large size, presents problems related to class imbalances, potentially resulting in biased model predictions if not adequately mitigated. Zhang et al. [5] stated that traditional CNNs will fail to address inter-class similarities and intra-class variations of dermoscopic images with poor classification accuracy. However, Shetty et al. [7] also reported that high-resolution feature extraction was computationally demanding and may make its real-time applications.

Strengths

The latest studies are characterised by applying cutting-edge techniques that set up transfer learning [6] so that the models exploit features learned from huge datasets in order to improve performances on limited medical data. Self-attention mechanisms as developed by Zhang et al. [5] enable the model to pay attention to specific parts of the lesion when making an accurate classification. Data augmentation strategies, as well as in Shetty et al. [7], are combined into one model, increasing the robustness of the model by exposing it to different lesion presentations .

Future Scope

Further studies could include abilities, thus beating the problems mentioned above:

- Data Augmentation and Synthesis: Producing synthetic images through GANs [2] for the rare classes of lesion can help curb this class imbalance.
- Hybrid Models: Combining CNNs and transformers based ensemble [4] may help in improving performance as transformers can capture global context efficiently.
- Explainability and Interpretability: Techniques such as saliency maps and SHAP values [5] will be very useful in the interpretation, thereby building trust with the clinicians.
- Mobile and Edge Deployments: Model optimization in edge devices may enable real-time diagnosis [7] in resource-restrained settings.

Such advancements in the future direction by researchers would further augment the clinical applicability of the skin lesion classification models.

In our research study, the comparative analysis of top advanced computer vision models has been done on the basis of the testing accuracy. Also we have done an explainable AI approach in our work that uses the application of GradCAM [8] technique on the sample model predictions during testing the performance of the model. GradCAM stands for gradient weighted class activation mapping and it is very popular in the explainable AI approach for visualization of computer vision model predictions.

METHODOLOGY

Dataset description

- The dataset used for this research was named "mammal88/skin_cancer" from hugging face which is actually the Skin Cancer MNIST: HAM10000 dataset [3]. It simply consists of images split into seven distinct classes, namely: Actinic Keratoses (akiec), Basal Cell Carcinoma (bcc), Benign Keratosis-like Lesions (bkl), Dermatofibroma (df), Melanoma (mel), Melanocytic Nevi (nv), and Vascular Lesions (vasc).

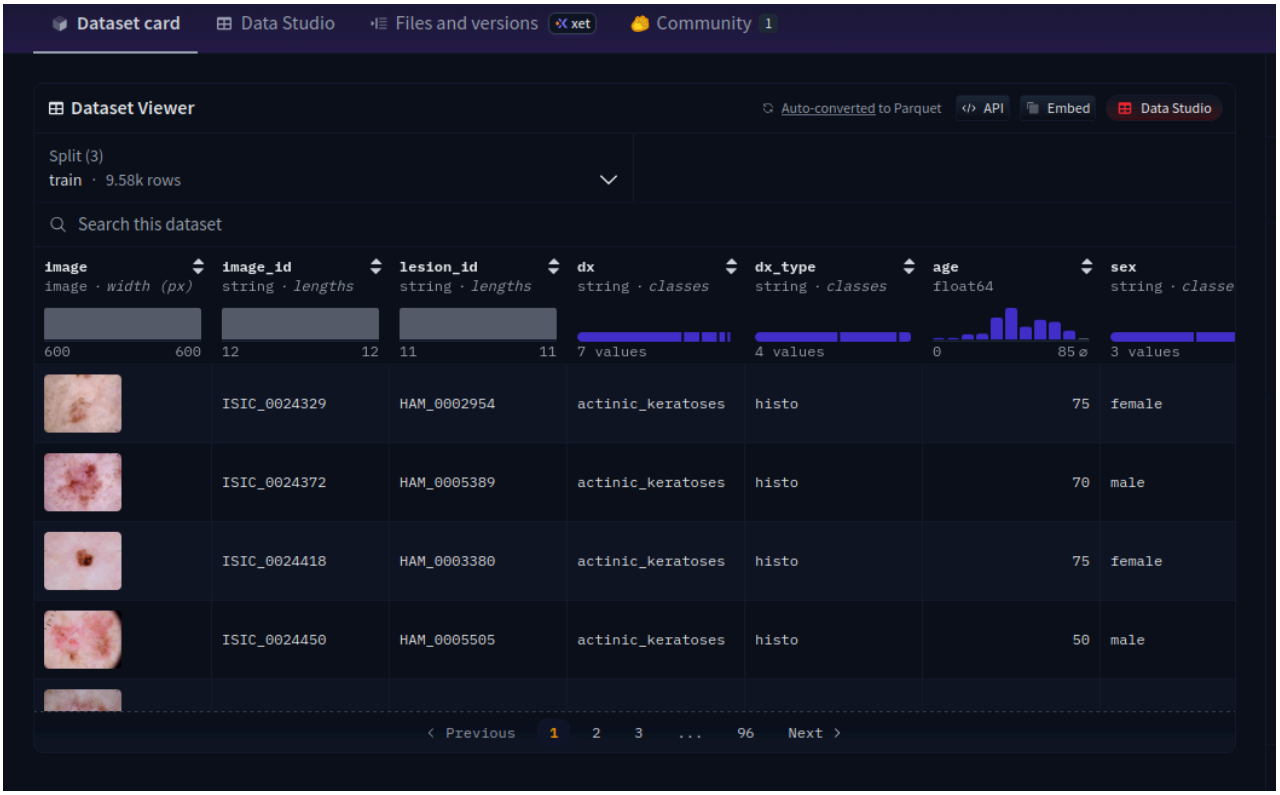


Figure 2: ham10000 dataset viewer in the hugging face

Below we have shared a few snippets of the skin cancer lesions from the dataset for the reference:

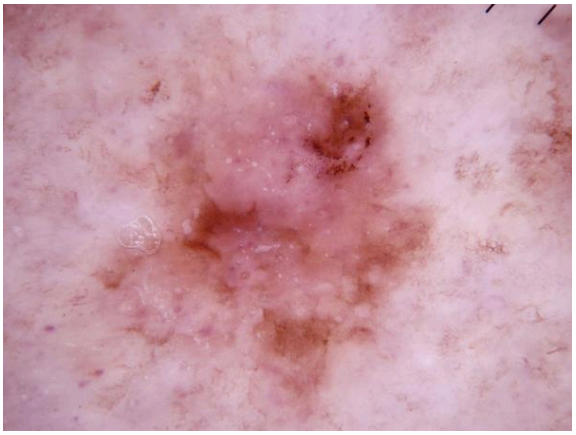


Figure 3 : actinic keratoses lesion condition image from the ham10000 dataset

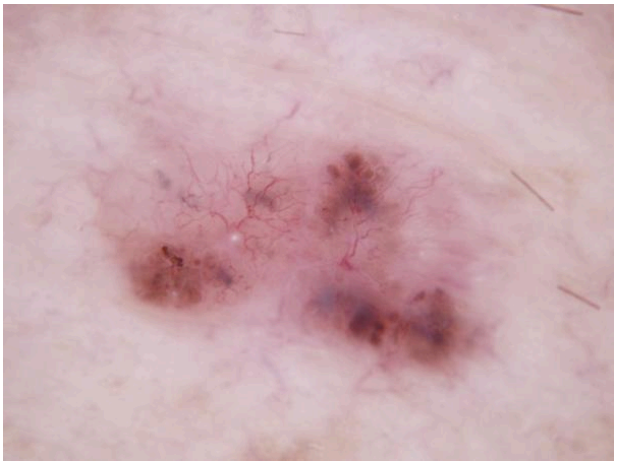


Figure 4: Basal cell carcinoma image from the ham10000 dataset

Figure 5 : Melanoma image from the ham10000 dataset

- Pre-split into training, validation, and test sets : this dataset is more systematically approached toward the construction and testing of models. Every image in the dataset contains metadata such as, image ID, lesion ID, dx, dx_type, age, sex, and localization.

Data Preprocessing Steps

- Resize the images: The sizes of the images are resized to 224x224 pixels to make them uniform in size and compatible with the model structure.
- Data augmentation approaches are also included to make the model more robust using random horizontal flips, rotation between ±20 degrees, and color jittering which includes change in brightness, contrast, saturation, and hue.
- Convert to PyTorch tensors.
- Normalization: Images are normalized using mean and standard deviations which can be loaded from the ImageNet dataset and very widely used for pre-trained models.
- Label Mapping: To train the model, the labels are mapped to numerical representations as dx. This mapping ensures that the label is interpreted and processed properly by the model.

Models used for training

We have listed the top computer vision models used on our dataset for the comparative study:

- FastViT_T8.apple_in1k : It is a hybrid vision transformer from apple [10] that uses structural reparameterization (RepMixer) to achieve low latency on the mobile devices and also outperforming pure cnn architectures like EfficienNet in terms of the speed-accuracy trade-off.

- Ghostnetv2_100.in1k: One of the popular models for efficiency and light architecture effectively suits any job, particularly for medical image classification.GhostNetV2 uses decoupled fully connected attention [9] in the low-cost “ghost” feature maps so that the model is able to capture the long range dependencies efficiently without sacrificing the speed and performance.

- EfficientViT_B0.r224_in1k: it uses a multi head convolution attention module [11] that maintains linear complexity while achieving much higher efficiency. It uses cascade-style architecture, combined with reparameterized convolutions and a sandwich layout of attention and feed-forward layers.

- ShuffleNet_V2_X1_0 : It is a highly efficient CNN that uses channel shuffle operations [12] and pointwise group convolutions to minimize computation while maintaining accuracy.

- LeViT_Conv_128s.fb_dist_in1k: a fast inference focused vision transformer [13] disguised as ConvNet

- MobileViT_XXS : an ultra lightweight hybrid model that replaces some convolutional blocks with mobile optimized transformers [14], delivering ViT like global reasoning with cnn level speed.

- MobileNetV4_CoNV_Small.e2400_r224_in1k : a compact pure cnn variant optimized for edge devices with universal blocks [15] for improved accuracy at low FLOPs.
- Resnetv2_50.a1h_in1k: it is the second generation variant [16] of the resnet 50 model and therefore it is better due to modern training techniques and pre-act.
- Resnext50_32x4d.a1h_in1k : cardinality enhanced model having grouped convolutions [17] and therefore it has far better efficient model architecture than resnet 50.
- Mobileone_s0.apple_in1k: Apple's efficient model having been trained on imagenet 1-k and having the re-parameterization approach [18] in its architecture for fast inference.
- Efficientformerv2_s0: It is a hybrid model consisting of efficient pure convolutions at early stages and lightweight attention mechanism at the later stages [19] to capture the local and global context in the images while training the model.

These computer vision models attain a very good balance between the computational efficiency and overall performance. These are pre-trained models on ImageNet dataset that give good feature extraction capabilities and therefore these models are fine-tuned on the skin lesions dataset.

Hyperparameters

Optimizer : Uses AdamW optimizer with learning rate = 1e-3 and weight decay = 1e-4. AdamW (Adam with Decoupled Weight Decay) [20] is an optimization algorithm used in the training of all the modern computer vision models for the ham10000 dataset. It improves the standard Adam by separating the weight decay step from the adaptive gradient update. AdamW applies the weight decay directly to the model parameters and that's why it is the most popular optimizer used in almost all the training of the deep learning models.

Learning rate Scheduler: it is a technique used in the training of computer vision models for dynamically adjusting the learning rate during the optimization process and therefore this helps the model to converge faster in early stages of training with a larger learning rate and then fine tune more precisely in smaller steps, leading to better stability and performance . Therefore in our research purpose, we have used different types of learning rate schedulers, which are as follows:

OneCycleLR scheduler will change the learning rate throughout training, ramping up to a peak pace i.e. high max_lr (warm-up phase) and then ramping down by gradually annealing to a tiny value while cycling momentum in the opposite direction. It follows the "1cycle" policy for super fast training and convergence.

ReduceLROnPlateau is the type of learning rate scheduler that dynamically adjusts the learning rate by monitoring the validation loss and therefore if the validation loss fails to improve then the scheduler reduces the learning rate which helps the model escape the plateaus in the early stages and fine-tune more effectively later, leading to better convergence of the model.

CosineAnnealingWarmRestarts is a learning rate scheduler that repeatedly applies cosine annealing followed by warm restarts. The periodic restart behaviour helps the optimizer escape poor local minima, explores the loss landscape more effectively and often leads to a better generalization than a plain decaying schedule, therefore making this scheduler a popular choice for computer vision training tasks.

Loss function

We have used two types of loss functions :
CrossEntrophyLoss is the de facto loss function used in the classification tasks and therefore it measures the difference between the predicted probability distribution and the true labels. It is defined by the formula. Therefore in our code, we use a weighted cross entropy loss function that minimises the class imbalance in the dataset. Class weights are calculated based on frequency occurring in every class in session of training.

FocalLoss is the advanced modification of the cross entropy loss function which is designed to handle extreme class imbalance and therefore this function is extremely helpful for medical datasets such as ham10000.

Training Loop

All the model trains for the 75 epochs. For every epoch, it will compute the training loss and training accuracy and evaluate the model on a validation set. The best model based on validation accuracy is saved.

Evaluation

The model is tested on the test set after training. Test loss, accuracy, and sometimes a confusion matrix showing the classification of the model across various classes are measured.

Explainable AI

We have applied the Grad-CAM method on the sample test images to visualize the model predictions and to understand how the model is behaving with respect to the local and global context in the sample test images.

IMPLEMENTATION AND RESULTS

Outlines of Findings are demonstrated in this section. We have successfully trained, validated and tested the computer vision models on the ham10000 dataset and therefore the results are quite surprising and promising for our research work. Below in the table, we have provided the results of our experiments of different models on the dataset:

Model	Batch size	Training accuracy	Testing accuracy
ghostnetv2_100.in1k	128	99.9478%	98.60%
efficientvit_b0.r224_in1k	256	99.9165%	98.29%
fastvit_t8.apple_in1k	128	99.9478%	97.90%
efficientformerv2_s0	16	99.3%	97.82%
shufflenet_v2_x1_0	256	99.1960%	97.43%
levit_conv_128s.fb_dist_in1k	256	99.9791%	97.43%
mobilevit_xxs	128	9925.86%	97.20%
resnetv2_50.a1h_in1k	16	95.85%	96.89%
mobilenetv4_conv_small.e2400_r224_in1k	256	99.6554%	96.81%
resnext50_32x4d.a1h_in1k	16	96.15%	96.19%
mobileone_s0.apple_in1k	32	94..69%	95.10%

Table 1: Summary of results of the models developed for skin lesion classification

Analysis & Discussion of the models performance

- GhostNetV2_100.in1k

Performance: The testing accuracy was the highest at 98.60%, while the training accuracy is about 99.9478%.

Why: GhostNetV2 is quite an efficient and lightweight architecture model, quite appropriate for tasks where great performance with minimal computational resources is required. Its architecture, which employs "ghost" modules in order to build more feature maps with fewer parameters, definitely contributed to achieving high performance in this challenge.

- EfficientViT_B0.r224_in1k

Performance: Test accuracy 98.29%, Training accuracy 99.9165%.

Why: EfficientViT combines good properties of ViTs and efficient design. It probably achieved its

excellent performance by the ability to capture global dependencies without an obvious decrease in computational requirements.

- FastViT_T8.apple_in1k

Performance: test accuracy is 97.90%, training accuracy is 99.9478%.
Why: This one is optimized for the sake of speed and efficiency. Hence, it may be the best suited for real-time applications. FastViT's performance is much worse compared to GhostNetV2, however competitive, probably because its priority is speed rather than precision accuracy.

- ShuffleNet_V2_X1_0

Performance: Testing accuracy of 97.43%, training accuracy of 99.1960%.
Why: ShuffleNet V2 was specifically optimized for mobile devices, notably with a channel shuffle and group convolutions. It's a little less accurate than the top models probably because of its focus on reducing the computational complexity rather than being tuned for maximizing accuracy.

- LeViT_Conv_128s.fb_dist_in1k

Performance: Testing accuracy of 97.43%, training accuracy of 99.9791%.
Why: LeViT is such an attempt to merge CNNs with ViTs to get the best from both worlds. In terms of its efficiency, LeViT has the same level of effectiveness as ShuffleNet V2, combining efficiency with accuracy.

- MobileViT_XXS

Performance: Test accuracy is 97.20% and the training accuracy is 99.2586%.
Why: Optimized for mobile and edge devices; developed models should be lightweight and efficient. Its performance is a notch inferior to other models, possibly because of focusing too much on bringing down model size and computing requirements.

- MobileNetV4_Conv_Small.e2400_r224_in1k

Performance: Test accuracy of 96.81% and Training accuracy of 99.6554%.
Why: Efficient convolution layers of MobileNetV4 make it very suitable for all mobile applications. Its performance is the weakest of all the models used here, probably because it has sacrificed accuracy to reduce computational complexity.

Comparison Analysis

- *Training vs. Testing Accuracy*: All models have high training accuracy, close to 100%, and also some models around 95 - 96 % and thus demonstrate the good learning of the training data. However, the accuracies in testing are different and GhostNetV2_100.in1k and EfficientViT_B0.r224_in1k scored the higher accuracy meaning good generalization towards unseen data.
- *Batch Size Effect*: Indeed, on average, the models with a larger batch size of 256 train better and contribute to greater training stability and better performance. However, exceptions apply: GhostNetV2_100.in1k performed outstandingly well with only 128.
- *Complexity of Model vs. Performance*: The more complex the model, then for example like GhostNetV2_100.in1k and EfficientViT_B0.r224_in1k, much higher accuracy is achieved. Less complex models like MobileViT_XXS and MobileNetV4_Conv_Small.e2400_r224_in1k are less accurate.

Confusion Matrix of all models on Testing Dataset

The confusion matrix has a more detailed view as to how the model is performing over the testing dataset, correct as well as wrong classification across classes. It gives a granular view of how the model is performing across the different classes in the testing dataset.

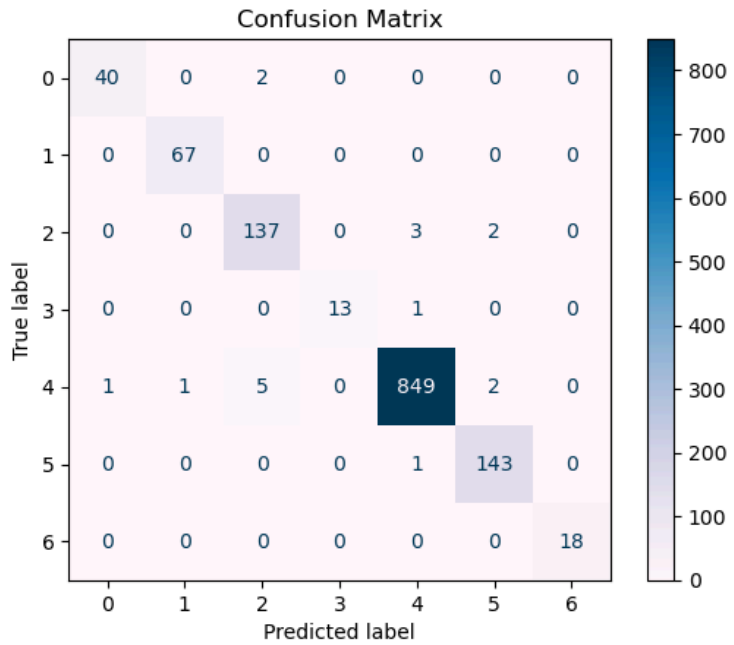


Figure 6. Confusion Matrix for GhostNetV2_100.in1k while the testing phase

Below is the granular analysis of confusion matrix generated for the GhostNetV2_100.in1k model:

- Class 0:

True Positive, TP: 40.
False Positives (FP): 0.
False Negatives (FN): 2.
Interpretation: The model correctly predicted 40 Class 0 instances, with two cases as Class 3. This indicates rather good accuracy and sensitivity for Class 0.

- Class 1:

TP: 67
FP: 0
FN: 0
Interpretation: The model correctly classified all the 67 Class 1 examples. That indeed translates to excellent precision and recall for Class 1.

- Class 2:

TP: 137
FP: 0
FN: 5
Interpretation: The model had correctly predicted 137 examples of Class 2, but assigned 5 to Class 4 and 2 to Class 5. Hence, it can be said that in Class 2, the model has a high precision and recall.

- Class 3:

TP: 13
FP: 0
FN: 1
Interpretation: It correctly classified 13 instances out of Class 3 and misclassified only one as Class 4, so that is good levels of precision and recall in Class 3.

- Class 4:

TP: 849
FP: 11
FN: 2
Interpretation: The model correctly classified 849 Class 4 events, failing to classify 11 as (1 Class 0, 1 Class 1, 5 Class 2, 4 Class 7). This confirms very high precision and recall in Class 4.

- Class 5:

TP: 143

FP: 0

FN: 1

Interpretation: The model correctly classified 143 instances of Class 5, which were incorrectly labeled as Class 4 only once. This relates to very high precision and recall values for Class 5.

- Class 6:

TP: 18

FP: 0 FN: 0

Interpretation: The model captured all 18 cases of Class 6, so it has really very good precision and recall as regards Class 6.

- *Precision and Recall*: The model demonstrates good precision and recall on classes 0, 1, 2, 3, 4, 5, and 6, that means the model is doing a good job in recognition of such classes.
- *Overall Performance*: With a testing accuracy of 98.60%, the GhostNetV2_100.in1k model performs extraordinarily well in the task of classifying skin lesions, misclassifying very few cases across classes.

Confusion matrix generated during the testing phase for the other computer vision models given below:

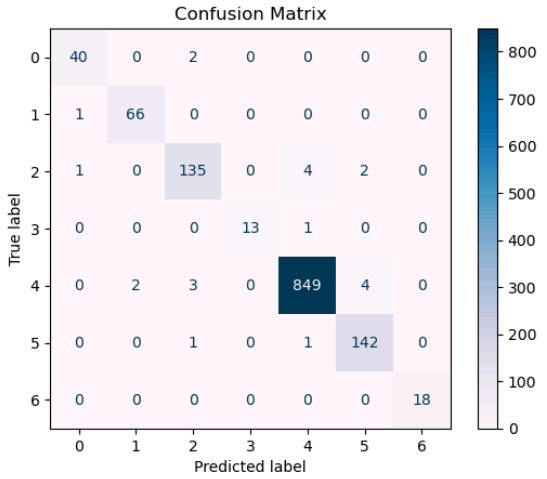


Figure 7. Confusion Matrix for EfficientViT_B0.r224_in1k

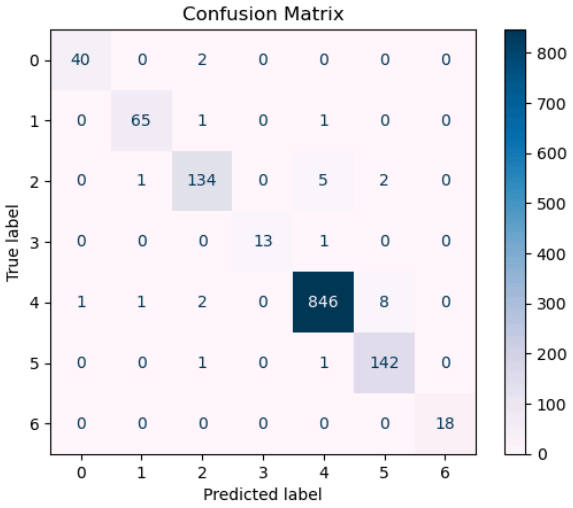


Figure 8. Confusion Matrix for fastvit_t8.apple_in1k

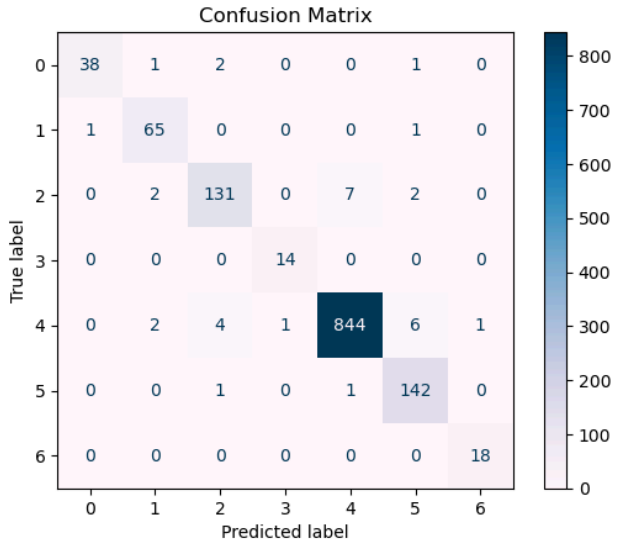


Figure 9. Confusion Matrix for LeViT_Conv_128s.fb_dist_in1k

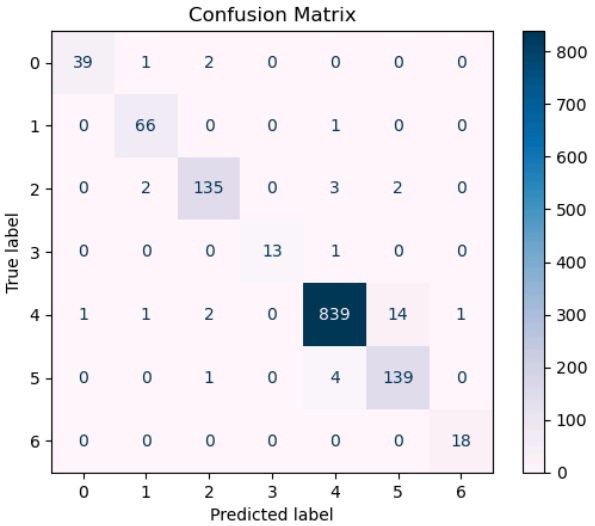
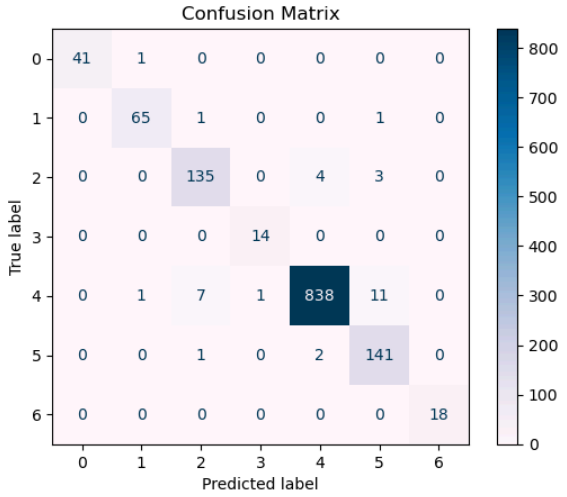


Figure 10. Confusion Matrix for MobileViT_XXS



■ Figure 11. Confusion Matrix for ShuffleNet_V2_X1_0

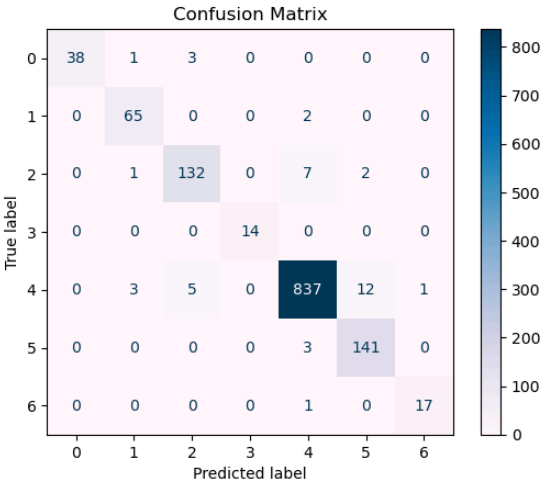


Figure 12. Confusion Matrix for MobileNetV4_Conv_Small.e2400_r224_in1k

Figure 13: efficientformerv2_s0 confusion matrix

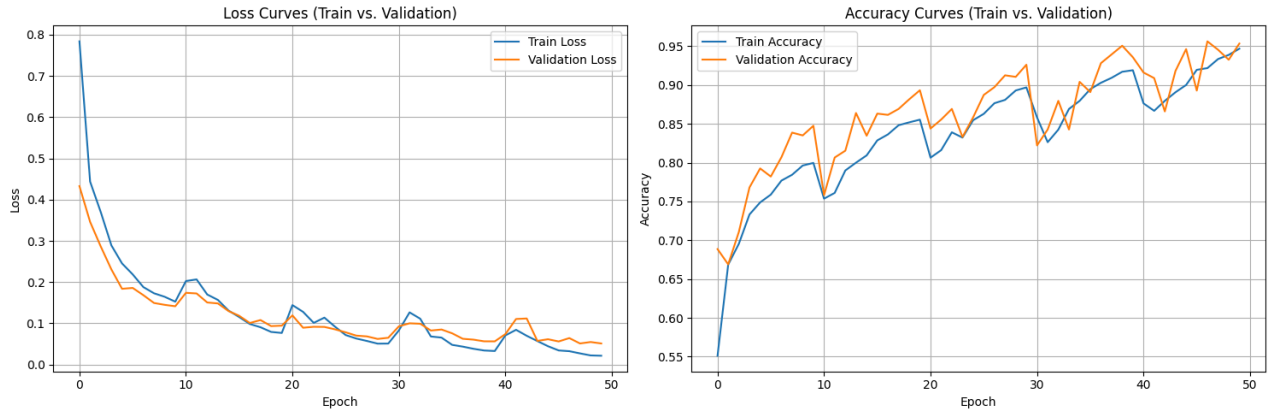


Figure 14. mobileone_s0 loss curves and accuracy curves

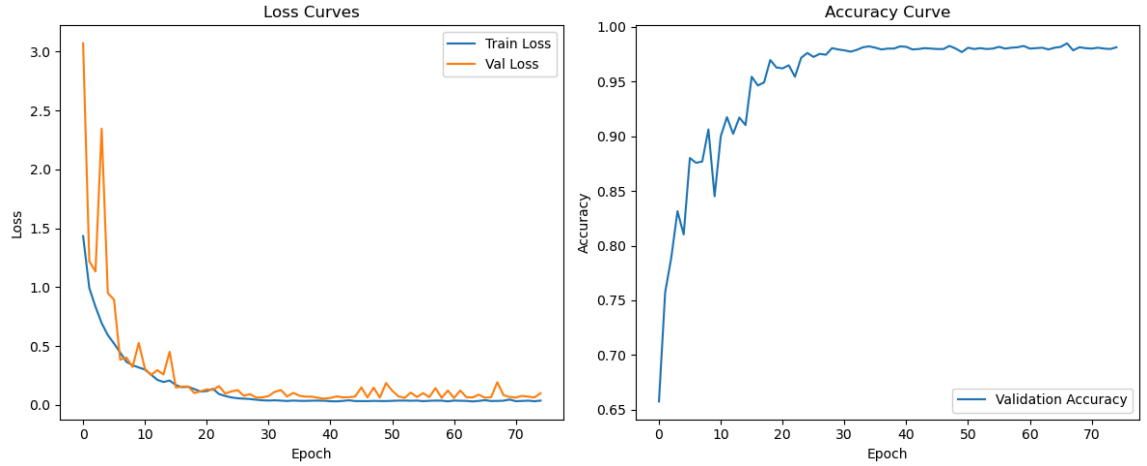


Figure 15. EfficientFormer loss and accuracy curves

In the figure 14, it is clearly evident that the mobileone model is performing very well on the training and testing dataset of ham10000 and the curves are going in a linear manner , like in a straight line.

Also in figure 15, it is very clear that the efficientformerv2 model is performing so well on the testing dataset of ham10000.

Therefore from the table 1 results , it is evident that these two models named as : GhostNetV2_100.in1k and EfficientViT_B0.r224_in1k, to be the better models for the classification of skin lesion tasks with good accuracy and reasonable generalization. Their architecture then turns out to be optimal for medical image classification, the point at which efficiency balances and meets enough accuracy. In such cases, the subsequent research would thus target further optimizations or even ensemble approaches to further push the performance.

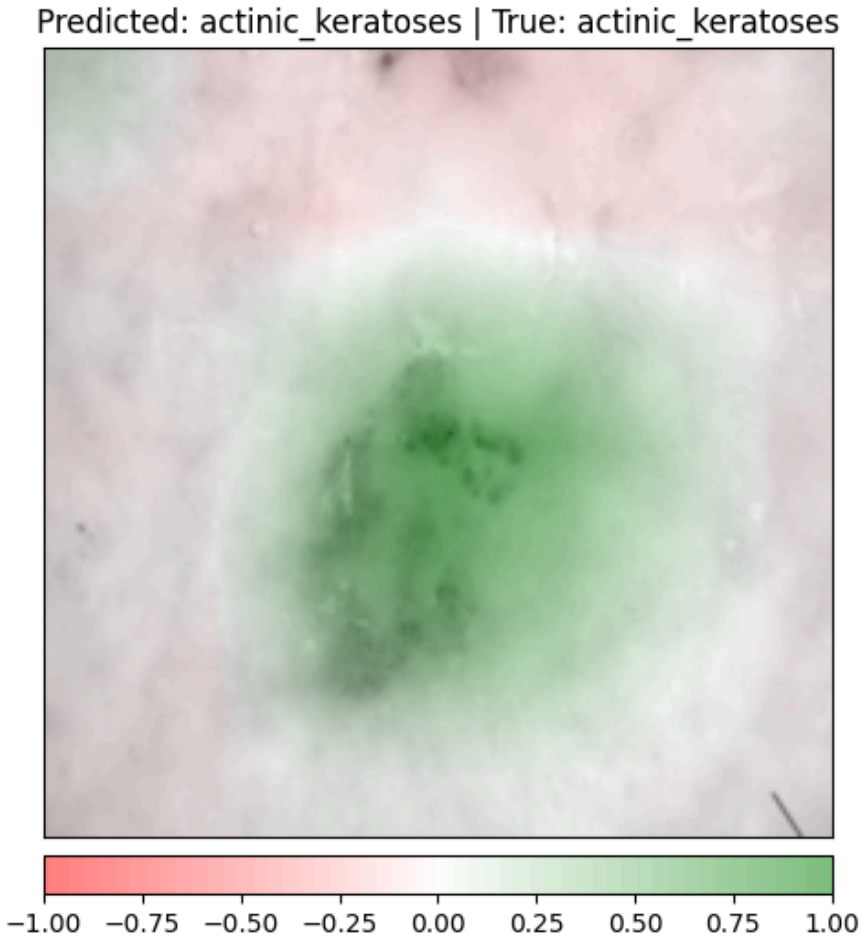


Figure 16: GradCAM of resnet50 on the sample test image of actinic_keratoses

From the figure 16, it can be assumed that the resnet50 is correctly behaving and it is focusing on the actual lesion on the skin for the prediction. Green areas in the gradcam suggest that the model thinks these areas are highly indicative of the skin lesion. Gradcam computes the gradients of the class score with respect to the last convolutional feature maps and therefore

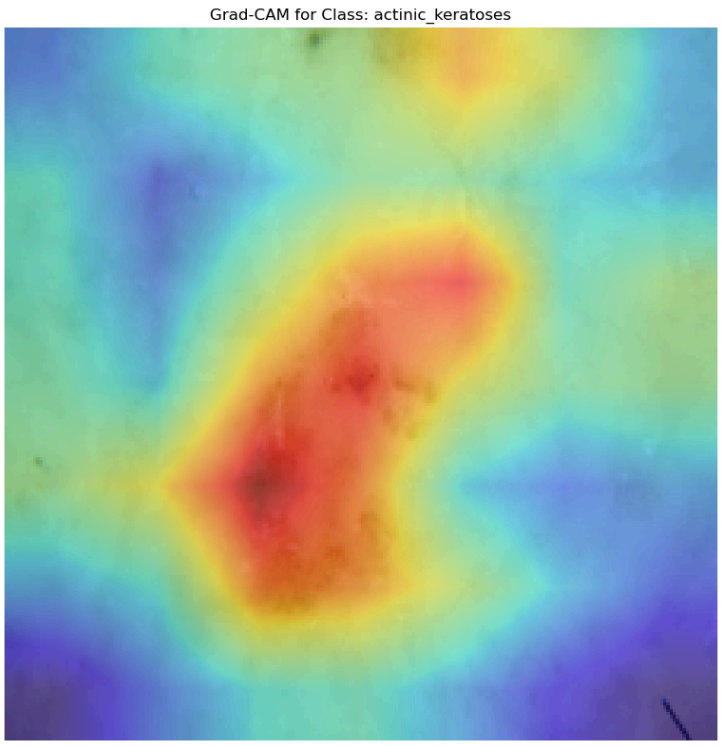


Figure 17. EfficientFormer Grad-CAM on the sample test image

Therefore from the grad-cam of efficientformer in the figure 17, it is clearly evident that the model is clearly focusing on the skin lesion area which is highlighted by the deep intense red color in the grad-cam for the skin lesion class : actinic keratoses, while making the prediction during the testing.

So we can assume and present that the grad-cam method is very useful in understanding how the computer vision model focuses on the patches of the sample test images to make the correct and precise class predictions and so explainable AI techniques build confidence and trust in the model predictions during the testing phase.

CONCLUSION

Overview of Work Completed

The research objective was the development and testing of deep learning models to classify skin lesions using the "mammal88/skin_cancer" dataset. The research was founded on large numbers of advanced models: GhostNetV2_100.in1k; EfficientViT_B0.r224_in1k; FastViT_T8.apple_in1k; ShuffleNet_V2_X1_0; LeViT_Conv_128s.fb_dist_in1k; MobileViT_XXS; MobileNetV4_Conv_Small.e2400_r224_in1k. The models were fine-tuned on the dataset, performing evaluations with the accuracy of training and testing.

The key findings include

- GhostNetV2_100.in1k reached the highest testing accuracy of 98.60%, demonstrating its efficiency for skin lesion classification.
- EfficientViT_B0.r224_in1k was again fantastic with a test accuracy of 98.29%.
- FastViT_T8.apple_in1k and others are competitive, and hence, it indicates that multiple architectures are valid for this task.
- The Confusion matrix analysis provided insights into how the model was performing on classes

Limitations of Work

- Generalization: While testing well, in fact, it is not possible to take it as applicable to real-life situations or other data sets. The validation must be performed with different kinds of data sets.

- Model Complexity: Some models such as GhostNetV2_100.in1k and EfficientViT_B0.r224_in1k are complex, hence computationally expensive that might limit its usage on resource-constrained devices.

- Data Augmentation: The study utilized simple data augmentation procedures. More sophisticated augmentation methods could help the model to perform better.

- Hyperparameter tuning: The models were fine-tuned using certain hyperparameters. Further optimization of these parameters may produce better results.

- Interpretability: Deep models, including transformer-based models, are perhaps less interpretable than their typical counterparts of machine learning models. Further insight into the decision-making process is likely to result from improvement of model interpretability. In short, deep learning models do come in handy in the classification of skin lesions; however, the high performers are GhostNetV2_100.in1k and EfficientViT_B0.r224_in1k. On the other hand, limitations and opportunities for improvement in deep learning models for real-world application certainly call for continued research and testing to ensure that the models so developed must be resistant and generalizable.

FUTURE SCOPE

Future Insights of Work

Skin lesion classification using deep learning models has produced excellent results, notably with models such as GhostNetV2_100.in1k and EfficientViT_B0.r224_in1k. However, there are various areas in which further research could improve the effectiveness and usefulness of these models.

1. Advanced Data Augmentation Techniques

- Generative Adversarial Networks (GANs): Using GANs to generate synthetic skin lesion images can help counterbalance the imbalance of classes and provide extra training data, thereby improving model performance.

- Transformations and Perturbations: The model learns better to generalize the novel unseen data by playing with more complex image transformations and perturbations.

2. Transfer Learning using Domain-specific Datasets

- Fine-tuning models on region-specific skin lesion datasets will enhance transfer learning by focusing on improving model performance in specific populations or geographic regions where the skin lesions may have different characteristics.

- Cross-domain transfer learning: The related domain is other dermatological disorders, so this will increase the flexibility of the model since it would classify a large number of skin diseases.

3. Ensemble and Hybrid Model Approaches

- Ensemble Models for High Accuracy: Combining the predictions of a series of high-performance models, such as GhostNetV2 and EfficientViT, can reduce the prospect of misclassifications, especially of difficult ones.

4. Interpretability and explanation

- Saliency Maps and Attention Visualization Techniques for Creating Visualizations Showing

That Model Attention to Parts of the Lesion Can Cast Light on What Would Actually Decide the Model that a Dermatologist May Understand and Be Convinced By Better Than Its Suggestion.

5. Real-time Implementation and Edge Computing:

- For the classification of skin lesions, optimization for deployment on mobile devices and edge computing platforms may make it more accessible in remote or underserved regions. It could offer real-time diagnostic support without consumption of excessive computational resources.

- Optimization for Low-Resource Environments: Further compression of the model size and reduction of the computational needs will confirm that the model remains functional on lower-powered devices without much loss in accuracy.

REFERENCES

1. A. Dosovitskiy et al., “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,” arXiv preprint arXiv:2010.11929, 2020.

2. T. G. Debelee, "Skin Lesion Classification and Detection Using Machine Learning Techniques: A Systematic Review," Diagnostics, vol. 13, no. 3147, pp. 1-40, Oct. 2023. doi: 10.3390/diagnostics13193147.

3. P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions," Scientific Data, vol. 5, no. 180161, pp. 1-13, Aug. 2018. doi: 10.1038/sdata.2018.161.

4. A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinger, "Skin Lesion Classification Using Hybrid Deep Neural Networks," Proc. of the International Conference on Computer Vision, pp. 1-5, Apr. 2019. Available: arXiv:1702.08434v2.

5. J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention Residual Learning for Skin Lesion Classification," IEEE Trans. Med. Imaging, vol. 38, no. 9, pp. 2092-2103, Sep. 2019. doi: 10.1109/TMI.2019.2893944.

6. A. Romero Lopez, X. Giro-i-Nieto, J. Burdick, and O. Marques, "Skin Lesion Classification from Dermoscopic Images Using Deep Learning Techniques," Proc. IEEE Biomedical Engineering Conference, pp. 1-4, Apr. 2017. Available: biomed-2017-paper.

7. B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhattacharya, and K. Lakshmanna, "Skin Lesion Classification of Dermoscopic Images Using Machine Learning and Convolutional Neural Network," Scientific Reports, vol. 12, no. 18134, pp. 1-12, Oct. 2022. doi: 10.1038/s41598-022-22644-9.

8. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, Oct. 2017, pp. 618-626.

9. Tang, Y., Han, K., Guo, J., Xu, C., & Wang, Y., “GhostNetV2: Enhance Cheap Operation with Long-Range Attention,” in Proc. NeurIPS 2022, 2022.

10. P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel and A. Ranjan, “FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

11. Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., & Yuan, Y., “EfficientViT: Memory Efficient Vision

Transformer with Cascaded Group Attention," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

12. Ma, N., Zhang, X., Zheng, H.-T., & Sun, J., "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," in Proc. 15th European Conference on Computer Vision (ECCV), Munich, Germany, Sept. 8-14, 2018, Lecture Notes in Computer Science vol. 11217, pp. 122-138, Springer, 2018.

13. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., & Douze, M., "LeViT: A Vision Transformer in ConvNet's Clothing for Faster Inference," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021, pp. 12259–12269.

14. Mehta, S. and Rastegari, M., "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," in Proc. Int. Conf. Learning Representations (ICLR), 2022.

15. D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, F. Yang, W. Wang, C. Ye, B. Akin, V. Aggarwal, T. Zhu and A. Howard, "MobileNetV4: Universal Models for the Mobile Ecosystem," in Proc. Eur. Conf. on Computer Vision (ECCV), 2024.

16. K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 630–638.

17. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 5987–5995.

18. P. K. A. Vasu, J. Gabriel, J. Zhu, Ö. Tuzel, and A. Ranjan, "MobileOne: An Improved One millisecond Mobile Backbone," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5318-5330.

19. Y. Li, J. Hu, Y. Wen, G. Evangelidis, K. Salahi, Y. Wang, S. Tulyakov and J. Ren, "Rethinking Vision Transformers for MobileNet Size and Speed," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16889-16900.

20. I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. Int. Conf. Learn. Representations (ICLR), 2019.