Machine Learning in Medicine Final Report

Multi-model Deep Learning Skin lesions Classification

Members

Trần Đăng An	*BI12-004*
Đặng Văn Tuấn	*BA11-096*
Nguyễn Anh Duy	*BI12-127*
Đoàn Hữu Thành	*BI12-418*
Vũ Đức Thành	*BI12-419*

January, 2024

Abstract

The increasing prevalence and complexity of skin diseases pose significant challenges in healthcare, making the development of artificial intelligence (AI) tools for skin disease diagnosis a pressing need. With the advent of the 4th industrial revolution and the application of AI in healthcare, there is potential to improve the quality of diagnosis and overcome limitations in resources and geographical barriers.The purpose of this study is to look into how AI and machine learning techniques can be used to help diagnose dermatological diseases. The study focuses on analyzing skin lesion imaging data and applying deep learning techniques such as feature extraction, network refinement, and model optimization. The research focuses on the classification of psoriasis, atopic dermatitis, and skin cancer. To improve diagnostic accuracy and efficiency, a variety of techniques are used, including image characterization, deep learning network refinement, and parameter optimization. The proposed methods are tested using both global open databases and skin lesion image databases. This study advances dermatological disease diagnosis and paves the way for better healthcare outcomes by leveraging the power of AI and machine learning.

Table of contents

Motivation

The skin is the largest part of the body with a large environmental contact surface, so the number of patients suffering from skin diseases has increased in recent years, and the disease types are very diverse and complex. According to a 2006 World Health Organization report, skin disease is the fourth leading cause of disease burden worldwide, affecting billions of people [1].

With the rapid development of the 4.0 revolution, artificial intelligence has been used very efficiently in healthcare around the world, including in Vietnam. Going to the doctor can be difficult, especially with the present epidemic. Many patients uploaded images (taken with personal phones) to doctors for guidance on diagnosis and therapy rather than visiting a doctor for a medical check. As a result, study and implementation of artificial intelligence (AI) technology to design tools to assist clinicians in increasing the quality of skin disease diagnosis is an essential demand, consistent with the 4th industrial revolution's development trend.

In AI, in addition to the necessity of data, machine learning techniques are an essential component of any successful AI project. A good machine-learning technique will help us get the most out of our data, allowing machine learning to be as effective as feasible. The more difficult the challenges to solve with AI, the greater the need for data and engineering. Dermatological disorders, in particular, are difficult to identify because of their small size, symptomatic nature, and numerous and complex presentations, especially when underlying diseases have been treated.

This research focuses on analytical, deep learning, and machine learning techniques. Technical resources automatically extract features from skin lesion imaging data, using thinning approaches and deep learning network modification, including :

Extraction techniques used to characterize photos of psoriasis, atopic dermatitis, and skin cancer.

Techniques to increase the quality of deep learning network models, such as sparsity and stratification

Deep learning network fine-tuning strategies include modifying, optimizing parameters, and connection designs.

Testing and assessing procedures using open datasets around the world and skin lesion picture databases.

I. DATASET

1. Overview

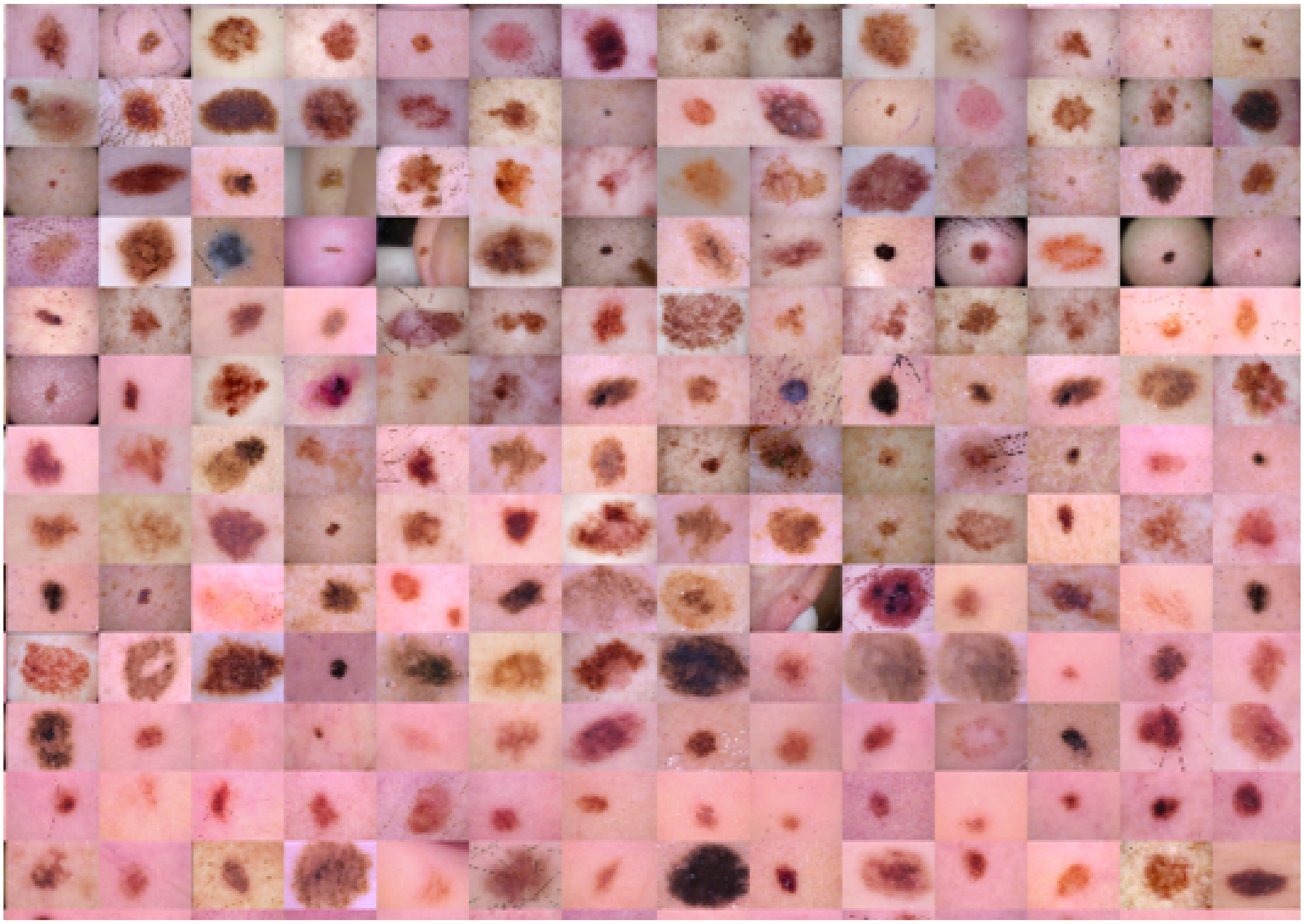

Fig. *A few HAM10000 pictures*

HAM10000 (Human Against Machine with 10000 training images) is a dataset of about 10,000 close-up, high-quality dermatological images of skin lesions for research on Artificial Intelligence for the problem of identifying dermatological diseases. The dataset was collected from the Department of Dermatology at the Medical University of Vienna, Austria, and the skin cancer clinic of Cliff Rosendahl in Queensland, Australia. The researchers used automation, machine learning, and manual evaluation to produce high-quality, consistent colors, middle lesions, and free-to-use. Specifics of the data collection process are found in their paper [2].

2. Component

The dataset consists of 7 disease labels denoted as follows:

- ak: Actinic Keratoses (Solar Keratoses) - Light Keratosis
- bcc: Basal cell carcinoma - Basal cell carcinoma
- bkl: Benign keratosis - Benign keratosis
- df: Dermatofibroma - Skin fibroids

- nv: Melanocytic nevus - Melanocyte microscope
- mel: Melanoma – Melanoma
- vasc: Vascular skin lesions

The images in the dataset were HAM10000 collected from patients presenting with skin-related symptoms and taken in high resolution. Each image is labeled with the corresponding type of skin pathology.

According to the author of HAM10000 [2], more than 50% of lesions are confirmed through histopathology, the actual label of the remaining cases is examination, monitoring (follow-up), expert consent (consensus), or in-vivo microscopy (confocal) identification. The data also includes lesions with multiple images (an area of damage taken with different images), which can be tracked with the accompanying metadata file. Not only can the lesion be observed and the corresponding disease label, but the metadata file also provides information about the method of identifying the disease (histo, follow-up, ...), the patient's gender, age, and the location of diseased skin.

The HAM10000 dataset has been used in many artificial intelligence and deep learning studies, aiming to develop machine learning models to detect and classify skin diseases. These models can be used to aid in the diagnosis and treatment of skin diseases. Google Scholar [1] contains about 2.5K results when searching for the keyword "HAM10000 dataset". The studies include:

- Karar Ali et al. [3] proposed a convolutional neural network model for predicting dermatological disease using HAM10000 datasets to evaluate the model.
- Garg et al. [4] proposed a convolutional network model for skin disease diagnosis with the HAM10000 dataset to train and evaluate the effectiveness of the model.
- Faes et al. [5] also proposed a convolutional neural network model for skin disease diagnosis using this dataset to evaluate the effectiveness of the model.

HAM10000 is offered to participants of the ISIC-2018 classification challenge[6] hosted by the annual MICCAI conference in Granada, Spain. In 2019, the ISIC-2019 challenge kicked off with a usage dataset that already included ISIC-2018 and added more new data. Information about it will be presented in a later section. In short, HAM10000 is the most common, foundational dataset for dermatology machine learning problems.

II. PROJECT’S OVERVIEW

1. Introduction

A technical summary of the research done to provide methods for characterizing data from skin lesion photos and network refining following model building is provided in this section. The dataset used for the study was HAM10000. The aim of this work is to use state-of-the-art machine-learning approaches to various datasets in order to increase the efficiency and accuracy of skin lesion analysis and diagnosis. Data on skin lesions was first gathered and preprocessed from the HAM10000 dataset. Numerous photos from various sources, such as internet archives and dermatology clinics, are included in each dataset. Professional dermatologists have carefully chosen and annotated the photos to guarantee precise labeling for teaching and assessment needs. Preprocessing methods like noise reduction, scaling, and normalizing are

1

2. Exploit characteristics

In order to extract features and patterns of information from pictures of skin lesions, feature extraction is a crucial step. Sophisticated methods are employed to automatically extract significant attributes. For this, pre-trained convolutional neural networks (CNNs) using massive picture datasets like ImageNet are frequently employed. As feature extractors, pre-trained CNN networks gather pertinent visual representations from intermediate layers in a triggered form. Subsequent classes or models then use these attributes for additional analysis.

3. Build models and refine

To categorize skin lesions, deep learning models are constructed once the trait has been extracted and thinned. For this, a variety of models constructed from deep neural networks (DNNs) are employed. First, each dataset's extracted features and basic labels are used to train a baseline model. However, fine-tuning approaches are used to maximize the model's performance. Refinement is the process of using a skin lesion dataset to update the parameters of a pre-trained network. This allows the model to adjust to the particular features and subtleties of each dataset, increasing accuracy and dependability.

4. Review

The generated methods are put through a rigorous evaluation and validation process with the use of relevant testing protocols and metrics. The HAM10000 dataset is split into subsets for training, validation, and testing in order to assess how well the suggested approaches perform. The model's performance on the dataset is measured using common assessment criteria like accuracy, accuracy, and dependability. Additionally, to guarantee that the model is generalizable across various dataset subsets, cross-validation techniques can be applied.

In conclusion, after developing models using the HAM10000 dataset, this technical study focuses on the research and development methods utilized to extract features from skin lesion imaging data and network fine-tuning. Skin damage analysis can be made much more accurate and efficient by utilizing several datasets and sophisticated deep learning techniques. The suggested methods should help doctors identify skin lesions correctly and aid in the early identification of skin conditions in a variety of datasets.

III. APPLIED DEEP LEARNING MODELS DETAILS

We have employed a range of sophisticated CNN architectures, including dense interconnect networks, networks, boot networks, and frameworks supporting search architectures, that have been developed in recent years. We applied these pre-trained models to ImageNet, a sizable dataset including over 1.5 million photos of natural scenes split into 1000 classes, in order to satisfy CNNs' insatiable appetite for data. Utilizing dermatology datasets, we improved these models to capitalize on the advantages of transfer learning. After experimenting with several CNN designs for this task, we ultimately selected ResNet-152[16], SE-ResNeXt-101[17], and VGG16 [30] because of their superior performance and wide range of model topologies.

1. ResNet-152

This section provides a thorough architecture of ResNet-152.A deep neural network specifically built for image analysis tasks, particularly in the context of error identification, is the ResNet-152 architecture. Stacked convolutional particles make up the central component of ResNet-152, which enables the network to effectively extract deep nonlinear properties from the input image. Stacking convolutional particles is how the convolutional process is accomplished in Figure 1. After a sequence of three-layer stacked convolutional transformations on the input image, rich nonlinear properties are produced in the form of tensor data. Because of pooled activity, typical neural networks lose local properties as depth increases. ResNet-152 uses residual connections to get around this problem.. These connections allow adding the input features of a certain layer to the output features, preserving important local information. It is important to note that the input and output features must be shaped the same for this shortening of identification to occur.

The ultra-deep architecture ResNet-152 has 152 convolution layers. The embedded three-layer stacked convolutional nucleus allows each convolutional layer to carry out three nonlinear transformations. Average aggregation is used in between each layer to reduce the feature size and increase the number of feature channels. This procedure aids in gathering and preserving pertinent data throughout the network. The 0.5 ratio Dropout function, which is coupled between each convolution layer to prevent over-equipment, randomly eliminates neurons, hence lowering their impact on the network's output. A dropout rate of 0.5 indicates that during training, half of the neurons are randomly eliminated on average. The number of output nodes in this fully linked layer will match the number of categories in the classification task when it is finally connected to the network.

The general architecture of ResNet-152 is shown in Figure, along with the information flow across the network. The network's capacity to comprehend complicated representations and attain high accuracy in image processing tasks is enhanced by the combination of stacked convolutional particles, redundant connections, and elimination regularization. The ResNet-152 design serves as an example of how well deep neural networks capture and use rich nonlinear characteristics for various image processing applications, including error diagnosis.

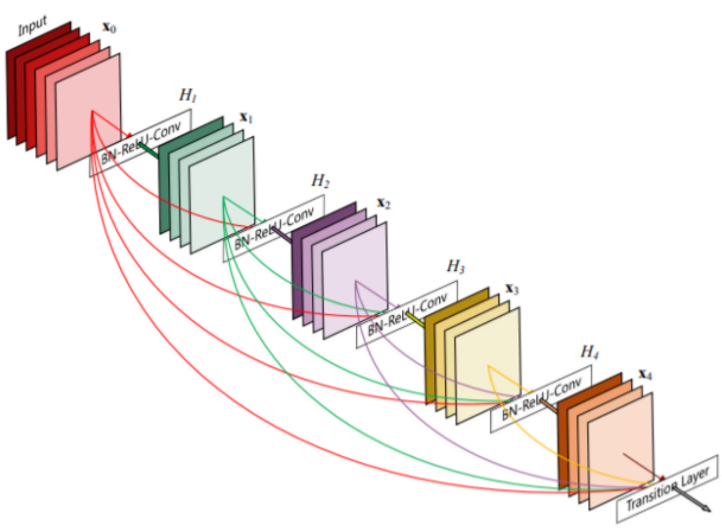

Fig. *The overall architecture of ResNet-152.*

We present the dynamics for reconstructing error characteristics and multiscale overlapping reception fields that may be incorporated into the design of deep redundancy networks. The convolutional nucleus that implements the local perception of the relevant input—a weighted sum over a local region of the input—is known as the receiving field. The convolutional nucleus that is most frequently utilized for feature extraction cannot be 1 because the convolutional nucleus must have a size greater than 1 in order to improve the cognitive field. .Even with a symmetrically added buffer (for instance, if the input is 4 × 4 and the convolutional particle size is 2 × 2, and the buffer on each side is 1, after sliding there will be a total of 5 outputs, which will not correspond to the input), an even-sized convolutional kernel cannot guarantee that the input feature map size and output feature map size do not change. Smaller convolutional particles contain more activation functions, richer features, and better clarity than bigger convolutional nuclei because more layers are stacked on top of one another. An activation function is associated with the convolutional activity, and the decision function can become more discriminating by employing more convolutional particles. Parametric calculation is required when swapping out a large-sized convolutional nucleus with a multilayer stacked convolutional nucleus. Table 1 compares the use of stacking for various network types (e.g., VGG-16, VGG-19, ResNet-50, and ResNet-152) and for the same type of network at different depths. Large-sized convolutional nuclei and multilayer convolutional nuclei have fewer parameters than multilayer stacked convolutional nuclei, as Table 1 illustrates, but the growth rate parameters remain steady at less than 1%, the replacement of 7 × 7 convolutional nuclei with 3×3+3 × 3+3 ×3 helix nuclei in the VGG-19 network has a minimum parameter growth rate of 0.06%, and the replacement of 5 × 5 convolutions nuclei with 3 × 3 + 3 × 3 stacked helix nuclei in the ResNet-152 network have the largest parameter growth rate of 0.9%.

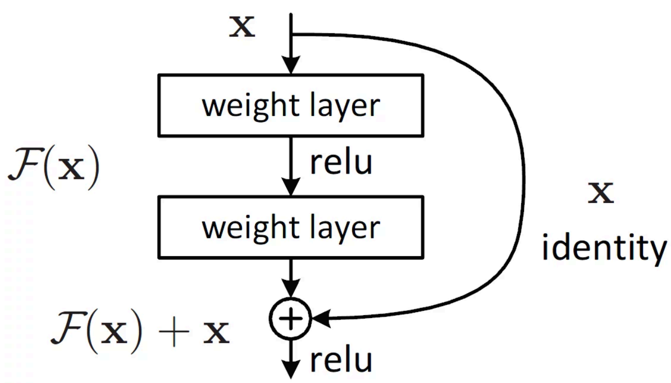

Fig. *The Residual Block architecture of ResNet-152.*

	3 × 3 + 3 × 3	**5 × 5**	**3 × 3 + 3 × 3 + 3 × 3**	**7 × 7**
VGG-16	18,952,131	18,310,787	19,137,027	19,102,595
VGG-19	53,595,203	53,544,835	53,779,715	53,746,051
ResNet-50	762,691	714,703	1,234,093	1,197,084
ResNet-152	2,419,171	2,397,061	6,679,779	6,434,577

Table. *Compare the number of computational parameters of multi-layer stacked convolutional particles inserted into different networks.*

2. SE-ResNeXt-101

A deep learning model called SE-ResNeXt-101 has shown impressive results in a number of computer vision tasks, such as picture categorization. It is an expansion of the ResNeXt design, which brings together the advantages of grouped convolutions and residual connections. Squeeze-and-Excitation (SE) blocks are a method that SE-ResNeXt-101 additionally presents. These blocks allow the model to adaptively adjust channel-wise feature responses. The integration of SE blocks with ResNeXt architecture has demonstrated significant efficacy in enhancing the representational capacity and discriminative capabilities of deep neural networks.

SE-ResNeXt-101 Model Architecture:

ResNeXt Framework: The ResNeXt architecture, which is predicated on the residual network (ResNet) idea, is expanded upon by SE-ResNeXt-101. Within each convolutional layer, ResNeXt adds a cardinality parameter that regulates the number of parallel routes or groups. Enhanced diversity and complexity of feature representations are made possible by this cardinality.

Blocks with Squeeze-and-Excitation (SE): SE blocks are incorporated into the architecture of SE-ResNeXt-101. Squeeze and excitation are the two stages that make up a SE block. The feature maps undergo global average pooling in the squeeze stage, which produces a channel-wise descriptor. Each channel's significance in relation to its total contribution to the feature representation is encapsulated in this description. In the excitation step, channel-wise dependencies are modeled and a set of channel-wise scaling factors is produced using a fully connected network with a sigmoid activation function.

Channel-Wise Recalibration: The model may adaptively recalibrate the channel-wise feature responses thanks to the SE blocks in SE-ResNeXt-101. By recalibrating, the network can now give distinct channels varying weights or priorities while concentrating on the most discriminative and informative characteristics. SE-ResNeXt-101 increases the network's capacity for representation learning and strengthens its discriminative potential by highlighting significant channels and suppressing less significant ones.

Training and Inference: Using massive picture datasets and supervised learning methods like stochastic gradient descent (SGD) or adaptive optimization algorithms like Adam, SE-ResNeXt-101 is taught. The model uses a sequence of convolutional layers, SE blocks, and pooling operations to extract hierarchical and context-aware features from an input image during inference. A classifier that maps the learnt features to the associated class probabilities—such as a fully connected layer or a softmax layer—is used to determine the final prediction.

Performance and Applications: SE-ResNeXt-101 has proven its ability to capture complex features and enhance the overall accuracy of deep learning models by achieving state-of-the-art performance on a variety of image classification benchmarks. Applications where accurate and reliable classification is crucial, like object recognition, scene comprehension, and medical picture analysis, have made extensive use of it.

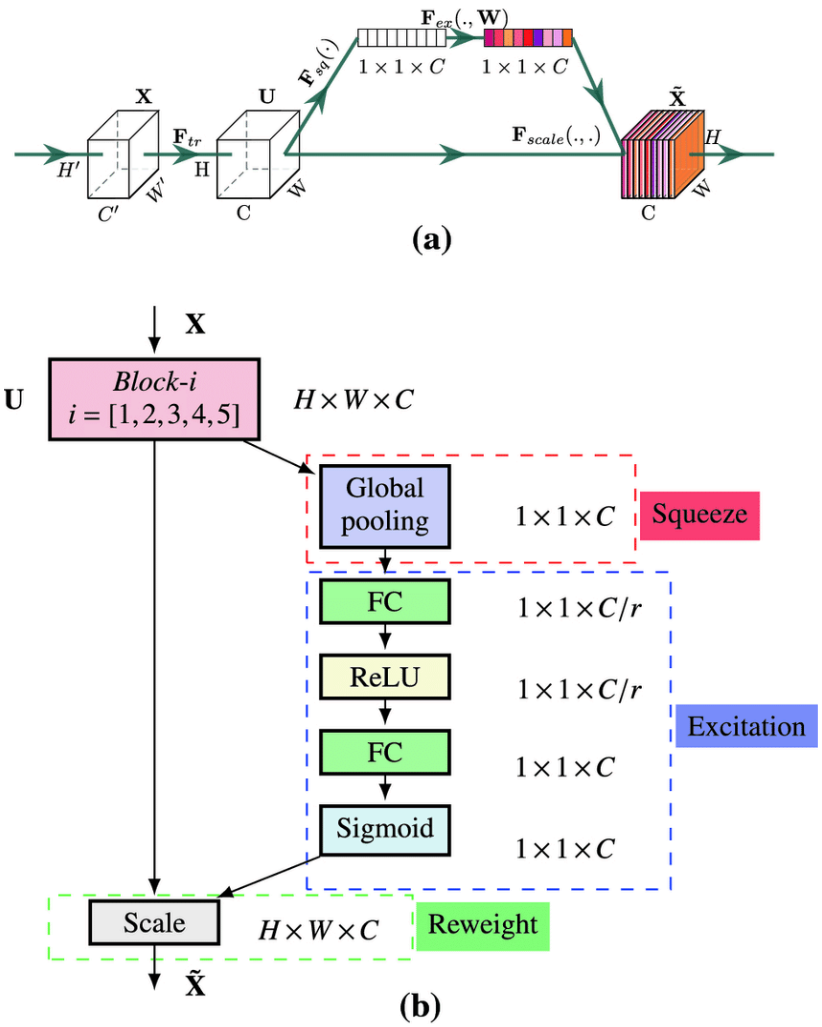

Fig. *The model's architecture of* SE-ResNeXt-101

The Squeeze-and-Excitation (SE) block consists of two key steps: squeeze and excitation. In the squeeze step, global average pooling is applied to the feature maps, which results in a channel-wise descriptor. Global average pooling computes the average of each channel across spatial locations, reducing the spatial dimensions to 1x1. The resulting descriptor captures the importance of each channel in terms of its contribution to the overall feature representation. In the excitation step, the channel-wise descriptor is fed into a fully connected network, typically comprising one or more fully connected layers followed by non-linear activation functions. This network aims to model channel-wise dependencies and generate a set of channel-wise scaling factors. These scaling factors represent the importance or relevance of each channel and are used to recalibrate the feature responses. The recalibration process allows the network to assign different weights or importance to different channels based on their importance for the task at hand. By emphasizing important channels and suppressing less relevant ones, the SE block enhances the discriminative power of the network. This adaptability enables the model to focus on the most informative and discriminative features, enhancing its representation learning capability.

The SE-ResNeXt-101 deep learning model, with its combination of ResNeXt architecture and SE blocks, represents a significant advancement in the field of computer vision. Its ability to adaptively recalibrate channel-wise feature responses has demonstrated superior performance in image classification tasks, making it a valuable tool for researchers and practitioners in the scientific community.

3. VGG16

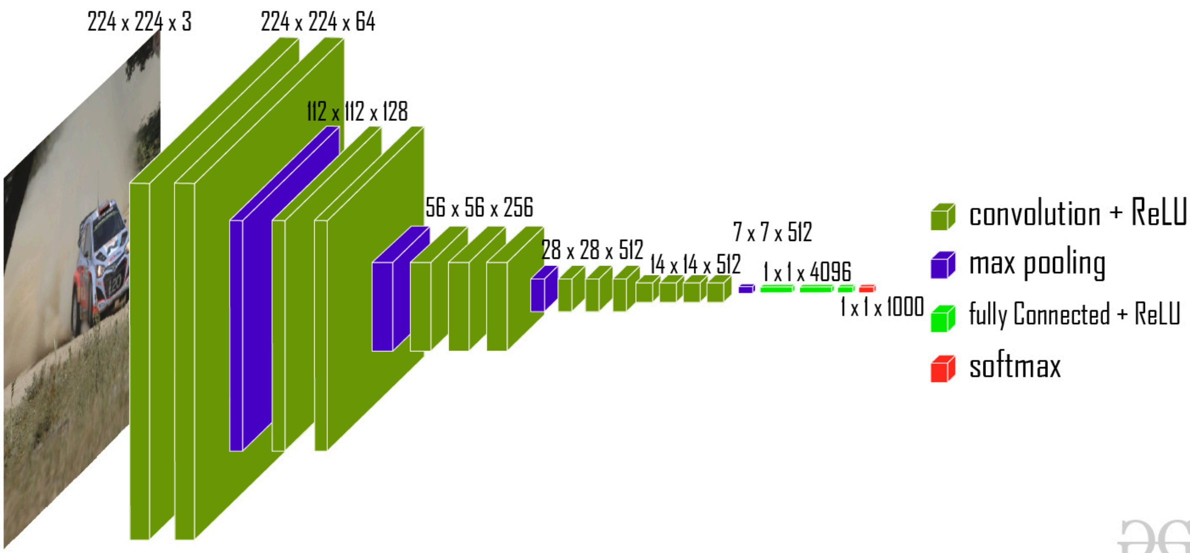

Fig. *Network structure diagram of VGG16*

The Visual Geometry Group (VGG) at the University of Oxford produced the convolutional neural network model VGG16, which was the 2014 ILSVRC object identification algorithm23 winner. One of VGG16's most important tasks is to show how, under some conditions, deepening the network might enhance its performance. The improvement of VGG16 over the original AlexNet can be attributed to the replacement of larger convolution cores (11×11, 7×7, and 5×5) with multiple 3×3 convolution cores. This can effectively broaden the depth of the network and improve its performance by reducing the number of network parameters. Thirteen convolutional layers, three fully connected layers, and five pooling layers make up the VGG16 network model.
VGG16 successfully enhances network performance while maintaining the basic features of the classical network and adding depth to it through flexible usage of 3 × 3 convolution. The VGG16 model does, however, have several limitations with regard to application. First, the VGG16 model faces challenges during front-end deployment due to the fully linked layer's numerous parameters, which take up a lot of memory and computational power. Second, compared to certain complex advanced networks, the network's performance is poor due to its single network model structure. Furthermore, VGG16 lacks a useful technique to stop gradients from disappearing, therefore issues like sluggish convergence and gradient explosion are likely to arise during model training. Aiming at the defects of the VGG16, this paper improved the VGG16 model by drawing on advanced network models such as ResNet, SqueezeNet, and DenseNet. The improved VGG16 model consists of 14 convolutional layers, five BN layers, six pooling layers, and one fully connected layer.

The network structures of VGG16 and Improved VGG16 are shown in Fig.
The following are the primary methods of improvement for the upgraded VGG16 model. Eliminate the VGG16 layers that are totally related to F6 and F7. Include the Global Average Pooling Layer (GAP) and Conv6. In the VGG16, the two completely linked layers FC6 and FC7 voluntarily link every

neuron to every other neuron in the layer above, producing a large number of parameters and using up a lot of processing power. Consequently, it is necessary to remove these two totally connected layers. The fully connected layer can be replaced with a novel concept called GAP, which was put forth by M. Lin et al. (2014). Experiments have shown that GAP can lower the number of parameters, computation load, and overfitting in the model24. In order to reduce the number of parameters, the quantity of calculation, and the over-fitting, GAP can determine the mean value of the pixel points in each feature map, output a feature point, fuse these feature points into feature vectors, and input them to the Sofmax layer. Additionally, GAP may produce a feature graph for every category, giving features a true meaning and creating a more logical connection between each category and feature graph. A GAP has been incorporated into numerous sophisticated network models, including GoogLeNet, ResNet, SqueezeNet, and DenseNet, as more and more studies have verified the operation of GAP. To equalize the size of the input and output channels, SqueezeNet25 introduced a convolutional layer before the GAP with a convolution kernel size of 1×1. This operation again reduced the number of parameters and computations in the model and significantly accelerated the speed. Therefore, in this paper, a convolutional layer Conv6 with a convolution kernel size of 1×1 was placed in front of the added GAP to optimize the model further. The number of filters on the model performance was analyzed by setting various filters (128/256/512) during model construction. [30]

IV. TECHNIQUES TO IMPROVE MODEL PERFORMANCE

1. Dropout function

In recent years, deep learning has emerged as a powerful paradigm in the field of artificial intelligence, revolutionizing various fields such as computer vision, natural language processing, and speech recognition. However, deep neural networks tend to be over-learned, making them unable to generalize well to unseen data. To address this challenge, the dropout function has attracted considerable attention and has become an integral component of deep learning architecture. We present a thorough analysis of the dropout function, its foundations, and its various deep learning applications in this paper. We explore the theoretical foundations of dropouts and elucidate their function in preventing overarming by removing units at random from both the training and inference stages. Additionally, we go over various approaches to dropout incorporation in various neural network topologies, such as transformer models, recurrent neural networks (RNNs), and convolutional neural networks (CNNs). We emphasize how dropouts reduce model complexity, improve regularization, improve model generalization, and lessen the effect of input noise.

2. Image Augmentation

In the fields of computer vision and image processing, picture augmentation is a commonly employed approach that is essential to improving machine learning model performance. To produce new, synthetic training data, it entails transforming preexisting photos in a number of ways. The enhanced photos differ in look, orientation, scale, and other aspects, but they still have the same semantic meaning as the originals. Because it may address problems like overfitting, lack of data, and difficulties with generalization, this method has become increasingly popular in scientific study, especially in the field of deep learning.

Image Augmentation Techniques:

- Rotation: A method of simulating various viewing angles and orientations by rotating an image to a particular degree.
- Translation is the process of moving a picture vertically or horizontally to mimic perspective or positional shifts.
- Scaling is the process of resizing an image to a larger or smaller size in order to simulate zooming in or out.
- Flipping: Mirroring a picture either vertically or horizontally so that models can gain knowledge from both positions.
- The process of adding random noise to an image to make it more resilient to noise in real-world situations is known as noise injection.
- Adjusting an image's brightness and contrast levels to take different lighting situations into consideration.
- Cropping is the process of removing unwanted areas from an image to concentrate the model's attention on certain characteristics.
- Gaussian Blur: Applying a blurring effect to an image, simulating out-of-focus or hazy environments.
- Color Channel Shifting: Altering the intensity or balance of color channels in an image, enabling models to handle variations in color composition.
- Elastic Distortion: Deforming an image using elastic transformations, introducing local spatial variations to the data.

Researchers can greatly increase the size of their training datasets and enhance the deep learning models' capacity for generalization by implementing these picture augmentation approaches. In the end, these enhanced images improve the overall performance and dependability of scientific research in computer vision and image analysis by providing diversity and variability that help models build robust and invariant representations.

3. Batch Normalization

Batch normalization is a popular deep learning technique that has helped convolutional neural networks (CNNs) perform well in a variety of computer vision tasks, such as picture categorization. The objective is to tackle the issue of internal covariate shift by standardizing the activations of every layer throughout a small subset of training cases. In doing so, batch normalization contributes to the stabilization and acceleration of the training process, improving convergence and deep learning model performance.

Batch Normalization Technique:

- Normalization: In batch normalization, each neural network layer's activations undergo a normalization process. It normalizes the activations to have zero mean and unit variance and computes the activations' mean and standard deviation within a mini-batch. The internal covariate shift—a term used to describe the movement in the layer input distribution during training—is lessened with the aid of this normalization step.
- Scale and Shift: Two learnable parameters, scale and shift, are introduced for each normalized activation following normalizing using batch normalization. With the help of these parameters, the model is able to determine the ideal scale and shift factors, which essentially permits the network to adjust while maintaining the representational capacity of the original network.

- Training and Inference: In order to normalize the activations, batch normalization computes the mini-batch's mean and standard deviation during training. By using backpropagation, the scale and shift parameters are updated. To maintain consistency between training and inference, the mean and standard deviation of the complete training set are utilized to normalize the activations during inference.
- Impact and Benefits: In deep learning, batch normalization has a number of advantages. By lessening the gradients' reliance on the parameter scale, it speeds up training and permits higher learning rates. Additionally, it lessens the issue of vanishing or inflating gradients, which facilitates the training of deeper networks. Furthermore, batch normalization serves as a kind of interlayer normalization, strengthening the network's resistance to variations in the distribution of input and enhancing generalization capabilities.

Researchers have seen significant gains in a range of picture classification tasks by integrating batch normalization into deep learning models. It is a key technique in deep learning and has advanced computer vision research because of its capacity to handle internal covariate shifts and stabilize the training process.

4. Fine-tuning model

Model refinement refers to the process of modifying and adapting a pre-trained deep learning model to a new task or domain. Refinement involves using a pre-trained model, which has been trained on a large dataset and further trained on a dataset dedicated to the smaller task.

Here's an overview of the concept of fine-tuning change and how to use it in deep learning:

- Pre-trained models: Deep learning models, especially large-scale models such as Transformative Neural Networks (CNNs) or Transformer models, are often pre-trained on large datasets, such as ImageNet or large text sets. These pre-trained models learn common features that can be transferred to different tasks or domains.
- Transferable learning: Refinement promotes transferable learning, in which knowledge gained from previous training is transferred to a new task. Instead of training deep learning models from scratch on a small dataset, pre-trained models are used as starting points and then refined to tailor them to a specific task or domain.
- Task-specific adaptation: During fine-tuning, the weights of the pre-trained model will be updated using a smaller task-specific dataset. The final classes or a subset of classes in the model are often replaced or modified to fit the target task. The model is then trained on the new dataset and the weights are updated to optimize performance for the specific target task.
- Benefits of Refinement: Changing fine-tuning offers several benefits in deep learning. It allows deep models to be trained efficiently even when the task-specific dataset is limited, as it leverages the knowledge learned from the pre-training stage. Refinement also leads to faster convergence and better generalization, as the pre-trained model that has learned common features can transfer to the new task.
- Application Tuning: Tuning is widely used across many different areas of deep learning. In computer vision, pre-trained CNN models are often fine-tuned for tasks such as image classification, object detection, or semantic segmentation. In natural language processing, pre-trained Transformer models such as BERT or GPT are

fine-tuned for tasks such as sentiment analysis, answering questions, or creating text. Refinements can also be applied to other areas such as speech recognition, recommendation systems, and reinforcement learning.

Refinement change provides a practical and effective way to apply deep learning to new tasks or areas by leveraging pre-trained models. It enables efficient use of limited data, faster convergence, and improved performance of the target task.

V. TESTING AND EVALUATION

1. Related work

For a long time, researchers have focused a lot of emphasis on automatic identification of skin diseases. Though several classes are accessible, the majority of these investigations are restricted to binary or tertiary classification. Given the growing risk that melanoma represents to patient survival with each day that goes by, it makes sense to emphasize the significance of early identification. Though they might not be as deadly as melanoma, thousands of other skin conditions can still significantly lower a patient's quality of life. Our findings show that DL is quite capable of teaching hundreds of classes at once. We think that now is the ideal moment to fully utilize Deep Learning and start carrying out significant research that will eventually lead to an automated skin disease diagnosis system that is accepted by the industry. These technologies have the potential to have a significant impact on society since they not only assist dermatologists in clinical settings with diagnosis but also offer low-income patients in developed and developing nations with affordable and efficient initial screening. The fact that many researchers employ public or private datasets with their own selection of training/test division (albeit randomly chosen) and layer count is another factor to take into account while applying DL in dermatology. Because of this, comparing various classification techniques often yields little to nothing in common, as mentioned by Brinker et al. [20]. A sizable, publicly accessible, standardized dataset with well-defined training/test separations and benchmarking performance measures can be gathered and maintained to address this incomparable issue. While some publicly available datasets (like the HAM10000 dataset) do provide this training/pre-test split, most of them are tiny in size and have goals that are restricted to binary or tertiary classification. Small dataset studies cannot be consistently generalized, and while the results can be published, they cannot serve as a foundation for the real-world diagnostics applications of AI. On the other hand, large public datasets often have a lot of noise, low-resolution images, or blurred stamps. Important useful information needed for a detailed classification of seemingly similar diseases will be lost in such low-resolution images or watermarks. In addition, non-visual metadata, such as medical history, is generally not available in medical imaging datasets. However, this additional information may be key to an accurate and definitive diagnosis. We were able to use a disease classification approach for the HAM10000 dataset and improve our outcomes by 2.5% (refer to Table A1). If multi-model datasets are managed and publicly available, AI can certainly leverage additional information to improve its classification performance. While understanding and interpreting the results of any AI-based classification, it is important to recognize that accuracy or even sensitivity and specificity may not describe the complete picture of a model's performance. For this reason, in addition to other performance indicators, the sub-receive zone's Performance Characteristics Curve (ROC) (AUC) is also provided. From an AI point of view, we can say that it is a challenging challenge to use low-resolution datasets and low-resolution watermarks to achieve an average sensitivity of roughly 80% with an average false positive rate of 1.6% (Table 3, Exp-2). acceptable level of success. However, as Navarrete-Dechent et al. pointed out, any AI-based classifier's real performance can differ greatly in a real-world clinical scenario. [22]. They found that the classifier developed by Han et al. [21] did not generalize well when presenting data from the archive on demographics differently than the data used to train the classifier. It's obviously reason for concern for a dermatologist. Han et al., however, argued in their response [21] that sensitivity and specificity alone should not be used to evaluate a classifier. ROC curves demonstrate the classifier's actual performance in generating diagnostic predictions for a given image below a number of operating points or thresholds. On the model's output, changing this threshold from 0 to 1 can alter the accuracy while also altering the balance between sensitivity and specificity. Therefore, a greater AUC value guarantees that the model can correctly forecast a specific disease, such as melanoma, with little probability of misclassifying any other disease as that specific disorder.

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)	AUC (%)
Resnet-152	60.82 ± 0.51	82.16 ± 0.43	98.50 ± 0.10
Densenet-161	63.51 ± 0.68	84.46 ± 0.46	98.49 ± 0.06
SE_ResNeXt-101	64.03 ± 0.77	84.26 ± 0.66	98.48 ± 0.08
NASNet	60.69 ± 0.72	81.09 ± 0.61	97.90 ± 0.03
Ensemble	66.74 ± 0.64	86.26 ± 0.54	98.77 ± 0.07

Table. *Detailed results of classification for individual classifiers and their aggregators on HAM10000.*

2. Our work and comparison

The SE-ResNeXt-101 model has a runtime of 4 hours and 30 minutes, which is equivalent to 270 minutes. Average runtime of 5.4 minutes per epoch. This indicates that this model has a relatively long average runtime per epoch.

The ResNet-152 and VGG16 models have a runtime of 2 hours and 30 minutes, which is equivalent to 150 minutes. An average runtime of 3 minutes per epoch. Compared to SE-ResNeXt-101, these two models have a shorter average runtime per epoch.

Model	Top-1 Accuracy (%)	Top-2 Accuracy (%)
Resnet-152	73.07 ± 0.02	71.69 ± 0.10
SE-ResNeXt-101	84.24 ± 0.18	82.50 ± 0.14
Vgg16	68.20 ± 0.22	68.70 ± 0.25

Table . *Detailed results of our classification for individual classifiers on HAM10000.*

1. ResNet-152:

· Top-1 Accuracy: 73.07%

- Top-2 Accuracy: 71.69%
- Validation accuracy: 73.07%
- Test accuracy: 71.69%
- Validation loss: 0.814018
- Test loss: 0.798522

2. SE-ResNeXt-101:

- Top-1 Accuracy: 84.24%
- Top-2 Accuracy: 82.50%
- Validation accuracy: 78.93%
- Test accuracy: 82.50%
- Validation loss: 0.6020
- Test loss: 0.4753

3. VGG16:

- Top-1 Accuracy: 68.20%
- Top-2 Accuracy: 68.70%
- Validation accuracy: 68.70%
- Test accuracy: 68.20%
- Validation loss: 0.942204
- Test loss: 0.961099

Based on these metrics, we can make the following observations:

- SE-ResNeXt-101 has the highest Top-1 and Top-2 accuracies and the lowest losses on both the validation and test sets. It has the best overall performance among the three models compared.
- ResNet-152 shows relatively stable accuracy and losses on both the validation and test sets, but its performance is lower compared to SE-ResNeXt-101.

VGG16 has the lowest accuracy and highest loss among the three models, indicating lower prediction capabilities compared to the other two models

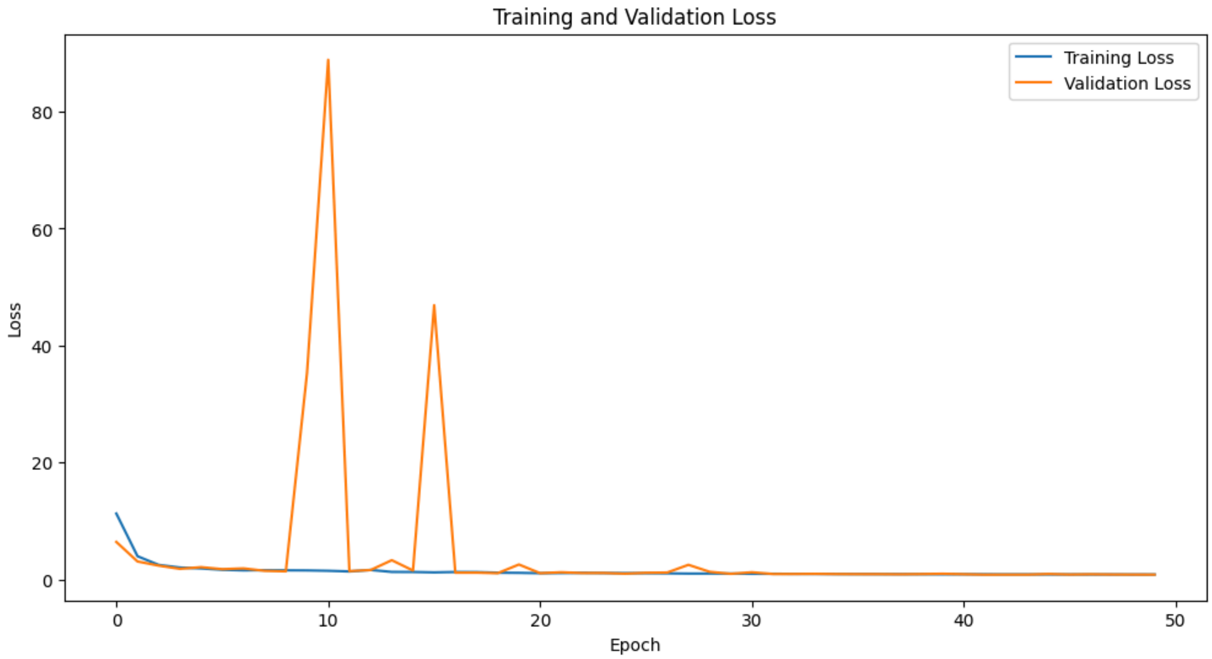

Fig. *The validation loss shows irregular fluctuations initially but gradually decreases towards the end, while the training loss continuously decreases throughout the training process. This pattern suggests the following observations:*

1. Early overfitting: The validation loss's early oscillations could be early warning signals of overfitting. Early in the training process, the model might pick up on certain traits from the training set, which would cause the training loss to steadily decline. Nevertheless, the model's performance might not generalize well when applied to the validation data, increasing the validation loss. This may occur if the training data is insufficiently representative of real-world data or if the model is very sophisticated.
2. Convergence: The gradual decrease in the validation loss suggests model convergence. This indicates that the model has learned and improved its predictive ability on the validation data. Despite the initial fluctuations, as the model continues to learn and adjust, it approaches a minimum loss value.
3. Performance evaluation: Despite the initial fluctuations, if the training loss consistently decreases and the validation loss gradually decreases towards the end, it indicates the positive performance of the model. Your model demonstrates the ability to learn from the training data and generalize well when applied to the validation data.

In conclusion, the observed pattern in the training and validation losses suggests that your model initially experiences overfitting but gradually converges and performs well on the validation data.

3. Discussion

Although this study used three different models, including ResNet-152, SE-ResNeXt-101, and VGG16, to classify skin inflammation images from the HAM10000 dataset, there are still some limitations that need to be considered to ensure the objectivity and accurate evaluation of our results.

- Sample size: The HAM10000 dataset is a large dataset consisting of 10,000 skin inflammation images from seven different disease types. However, using a fixed sample may lead to some limitations. This dataset may not fully reflect the diversity and variations of skin inflammation in reality. Therefore, the application of the results of this study needs to be carefully considered, especially for rare cases or cases that are not within the scope of the dataset.
- Reliability of diagnosis: In this study, the diagnosis of skin inflammation is determined based on the image classification process using three models: ResNet-152, SE-ResNeXt-101, and VGG16. Although all three models achieved relatively high accuracy, the diagnosis of skin inflammation can still have limitations. There may be cases where the classification process cannot accurately differentiate between similar disease types or there may be errors in classifying specific cases. These limitations can affect the reliability of the diagnosis results and the practical application of the method.
- Modeling and algorithm limitations: Although this study used three different models, all three have their own limitations. These models include ResNet-152, SE-ResNeXt-101, and VGG16, and each model has its own characteristics and limitations. The selection and tuning of hyperparameters, performance evaluation, and data processing can also affect the accuracy and stability of the results. The limitations in modeling can affect the accuracy and stability of the results and should be considered when applying this method in practice.

Potential development directions:

Although this study has achieved some notable results in classifying skin inflammation images using three models, ResNet-152, SE-ResNeXt-101, and VGG16, there are still many potential development directions to further research and improve our results. Below are some important development directions that we consider:

- Expand the dataset: Although the HAM10000 dataset provided a good platform for research, expanding the dataset by adding images from multiple sources and different hospitals can help increase the representativeness and diversity of skin inflammation. This will help improve the diagnostic ability and practical application of the method.
- Model optimization: Although the three models have shown relatively good results, fine-tuning and optimizing the hyperparameters and model structures can improve their performance. Further research can focus on exploring more powerful optimization methods to achieve higher accuracy and better generalization ability.
- Integration with other methods: The research can explore the possibility of combining deep learning models with other methods such as image segmentation or attribute classification to improve the diagnosis of skin inflammation. Combining information from multiple sources and methods can increase accuracy and provide additional information about the pathology and attributes of the skin.
- Building real-world applications: Finally, an important potential development direction is to build real-world applications based on the results of this study. Developing mobile applications or online tools can provide diagnostic support for doctors and end-users while collecting real-world data to improve the model and evaluate its performance in a real-world environment.

VI. Our project’s drawbacks

Regretfully, we had to use Google Colab to complete our project because of resource constraints. This limited our capacity to investigate more deeply and find more matrices to assess the model's effectiveness. Unfortunately, we were unable to advance in our attempt to obtain a more thorough comprehension of the model's potential. The limited computing resources made available by Google Colab made it difficult for us to conduct an analysis that went beyond the original parameters. As a result, we were unable to investigate sophisticated assessment techniques and minute details that would have offered insightful information about the model's advantages and disadvantages.
Although Google Colab provided accessibility and convenience, it also came with serious memory and processing power constraints. This limited our ability to do in-depth experiments and fully explore a larger variety of evaluation matrices, which would have improved our comprehension of the behavior and performance of the model in a number of different contexts. Despite these limitations, we made an effort to utilize all of the resources at our disposal and carried out a thorough investigation within the constraints set forth. It is imperative to recognize, though, that the results of our project must be viewed in light of the limited circumstances surrounding its execution.

In order to further investigate evaluation matrices and gain a better knowledge of the model's performance, it would be advantageous to look into pathways that offer more substantial computational resources. This would offer a more thorough and comprehensive evaluation, opening the door for further developments in our research projects.

VI. Bibliography

[1]	PE and LeBoit*, World Health Organi-zation Classification of Tumours: Pathology and Gienetics of Skin Tumours,* IARCPress, 2006.
[2]	P. Tschandl, C. Rosendahl and a. H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data,* vol. 5, 2018.
[3]	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," *Neuroscience Informatics,* vol. 2, 2022.
[4]	R. Garg, S. Maheshwari and A. Shukla, Decision Support System for Detection and Classification of Skin Cancer Using CNN, Singapore: Springer Singapore, 2021.
[5]	L. Faes, S. K. Wagner, D. J. Fu, X. Liu, E. Korot, J. R. Ledsam, T. Back, R. Chopra, N. Pontikos, C. Kern, G. Moraes, M. K. Schmid, D. Sim and K. Balaska, "Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study," *The Lancet Digital Health,* pp. e232-e242, 2019.
[6]	N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC),," 2019.
[7]	M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "Bcn20000: Dermoscopic lesions in the wild," 2019.
[8]	N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," 2017.
[9]	M. A. Kassem, K. M. Hosny, and M. M. M. Fouad, "Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning,," *IEEE Access,* vol. 8, 2020.
[10]	B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and a. M. H. Yap, "Analysis of the isic image datasets: Usage, benchmarks and recommendations,," *Medical Image Analysis,* vol. 75, 2022.
[11]	X. Y. a. W. L. A. Gong, "Dermoscopy image classification based on stylegans and decision fusion," *IEEE Access,* vol. 8, 2020.
[12]	T. Shanthi, R. Sabeenian, and R. Anand, "Automatic diagnosis of skin diseases using convolution neural network," *Microprocessors and Microsystems,* vol. 76, 2020.
[13]	M. N. Bajwa, K. Muta, M. I. Malik, S. A. Siddiqui, S. A. Braun, B. Homey, A. Dengel, and S. Ahmed, "Computer-aided diagnosis of skin diseases using deep neural networks," *Applied Sciences,* vol. 10, 2020.
[14]	T. A. Rimi, N. Sultana and M. F. A. Foysal, "Derm-nn: Skin diseases detection using convolutional neural network," *4th International Conference on Intelligent Computing and Control Systems (ICICCS),* 2020.

[15] P. Tschand, 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, 2018.

[16] Yu, H.; Miao, X.; Wang, H. Bearing Fault Reconstruction Diagnosis Method Based on ResNet-152 with Multi-Scale Stacked Receptive Field. Sensors 2022

[17] Ramadan, M.K.; Youssif,A.A.A.; El-Behaidy, W.H. Detection and Classification of Human-Carrying Baggage Using DenseNet-161 and Fit One Cycle. Big Data Cogn. Comput. 2022

[18] Jianxiang Dong. 2020. Focal Loss improves the Model Performance on Multi-Label Image Classification with Imbalanced Data. In proceedings of the 2nd International Conference on Industrial Control Network And System Engineering Research (ICNSER2020). Association for Computing Machinery, New York, NY, USA, 18–21.

[19] L. G. Falconí, M. Pérez and W. G. Aguilar, "Transfer learning in Breast mammogram Abnormalities Classification with Mobilenet and Nasnet," *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Osijek, Croatia, 2019, pp. 109-114

[20] Brinker, T.J.; Hekler, A.; Utikal, J.S.; Grabe, N.; Schadendorf, D.; Klode, J.; Berking, C.; Steeb, T.; Enk, A.H.;
von Kalle, C. Skin cancer classification using convolutional neural networks: Systematic review. J. Med

[21] Han, S.S.; Lim, W.; Kim, M.S.; Park, I.; Park, G.H.; Chang, S.E. Interpretation of the Outputs of a Deep
Learning Model Trained with a Skin Cancer Dataset. J. Invest. Dermatol. 2018, 138, 2275

[22] Navarrete-Dechent, C.; Dusza, S.W.; Liopyris, K.; Marghoob, A.A.; Halpern, A.C.; Marchetti, M.A.
Automated dermatological diagnosis: Hype or reality? J. Invest. Dermatol. 2018, 138, 2277–2279.

[30] A novel method for peanut variety identification and classification by Improved VGG16 HaoyanYang1 , Jiangong Ni2 , JiyueGao2 , Zhongzhi Han2* & Tao Luan1*