WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n2.12
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Image Classification of Skin Cancer Using Deep Neural Networks with Scaling Laws

Rongman Xu

Beijing No.171 High School, Beijing 100013, China

ABSTRACT

Skin cancer image classification is critical to improve healthcare outcomes. Current practice often involves time-consuming procedures that may delay diagnosis until the disease has progressed to an advanced stage, reducing the chances of successful treatment. This challenge is further exacerbated by the worldwide shortage of skilled dermatologists. In this study, we investigate the effect of dataset size on the image classification performance of eight networks (AlexNet, ResNet18, ResNet34, ResNet50, ResNet101, ResNet152, ViT, and MLP-Mixer). We trained these classifiers using different ratios (e.g. 1% to 100%) of samples from the HAM10000 dataset. Our experiment reveals the complex interplay among dataset size, model complexity, and skin cancer classification performance, validating the rules of the neural scaling laws on skin cancer image classification. This work highlights the impact of dataset scale and model complexity on improving the skin cancer image classification performance to potentially reduce the burden on healthcare professionals.

KEYWORDS

Deep neural network; Image classification; HAM10000; Skin cancer; Neural scaling laws

1. INTRODUCTION

Skin cancer is a significant health concern worldwide [1][2], and early detection and accurate diagnosis are essential for effective treatment and improved patient prognosis [3]. Dermoscopy [4], a non-invasive imaging technique, has become a valuable tool for the diagnosis of skin cancer, but manual interpretation of dermoscopy images by dermatologists is time-consuming and subjective [5]. For example, a research in 2019 suggested that the sensitivity and specificity of lesion classification by dermatologists were 67.2% and 62.2%, respectively, and in comparison, a pretrained ResNet50 had a sensitivity of 82.3% and a specificity of 77.9% [6]. Hence automated skin cancer classification methods are needed to overcome these challenges. Several convolutional neural networks (CNNs) [7][8] have been used to achieve significant results in skin cancer classification. Pomponiu et al [7] utilized 399 images to differentiate between melanomas and benign nevi. They applied data augmentation and preprocessing techniques, followed by utilizing a pretrained AlexNet to extract representational features. Subsequently, a k-nearest-neighbor classifier was employed using cosine distance metrics to classify the lesions. The algorithm underwent cross-validation, without testing on an independent dataset. It achieved a sensitivity of 92.1%, a specificity of 95.18%, and an accuracy of 93.64%. In Nasr-Esfahani et al [8], a two-layer CNN was trained from scratch to distinguish between melanoma and benign nevi based on clinical images. The training dataset consisted of only 136 images, and the test dataset contained 34 images, all sourced from the public image archive of the Department of Dermatology of the University Medical Center Groningen [9]. The method achieved a sensitivity of about 81%, a specificity of about 80%, and an accuracy of about 81%.

On the other hand, the amount of data required to achieve saturated accuracy with different dataset sizes and different model structures is still an unanswered question [10][11]. In real-world scenarios, hospital data is often region-specific and limited. Mainstream models may suffer from performance degradation upon analyzing such datasets [12]. To offer medical professionals comparative insights and judgments across various scenarios, in this work, a comprehensive analysis was conducted on the consistency of model structure and data scale according to the neural scaling laws in language models [13]. Specifically, this article evaluate the performance of various mainstream deep neural networks, including AlexNet [14], ResNets [15], ViT [16], and MLP-Mixer [17], on the classification of skin cancer images. The experimental results on the subsets of HAM10000 dataset [18] in different scales demonstrate that, both dataset size and model complexity play crucial roles on the accuracy of skin cancer classification models, exerting a significant impact on model performance. The goal is to promote the skin cancer detection that will ultimately improve healthcare outcomes and reduce the healthcare burden.

2. RELATED WORK

2.1. Deep Learning for Image Classification

In recent years, deep learning has greatly advanced image classification, allowing computers to understand images with unprecedented accuracy [19][20]. Deep learning allows computers to learn from large amounts of data and recognize patterns and features in images that would be too complex for traditional algorithms [21]. With this technology, computers can now recognize faces, diagnose diseases through medical scans, and even interpret satellite imagery with an accuracy that would have been unimaginable just a few decades ago [19][20][22][23][24][25].

CNNs are tailored to handle image data and automatically learn discriminative features. Unlike traditional methods like SVM [26], decision tree [27] and random forest [28] that required hand-crafted features, CNNs work by passing images through multiple layers of convolutional filters, gradually extracting and fusing features inspired by the human brain's visual processing system. This hierarchical feature extraction and combination enable CNNs to capture higher-level semantic information, making them effective for image classification and recognition [29][30].

There have been several important milestones in the development of CNNs in the field of image classification. One of these milestones is the AlexNet by Krizhevsky et al. in 2012 [14], which significantly reduced the error rate to 16.4% by winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by a large margin and demonstrated the power of deep learning in image classification. GoogLeNet [31] put together multiple convolution or pooling operations to form a network module that is designed to assemble the entire network structure in modules, reaching an error rate of 6.67% on ImageNet in 2014. In 2015, ResNet [15] had dropped the error rate to 3.57% by developing a residual connection technique, which greatly increased the depth of training deep neural networks. The emergence of these models will continue to drive the development of deep learning techniques for image classification.

In addition, significant advances have been made in network training Adam [32] techniques and data augmentation [33] strategies to improve the performance of deep neural networks for image classification. What’s more, transfer learning has been effective on reducing the required amount of annotated training data and improving the models’ generalization ability [34].

Overall, deep learning techniques have revolutionized the field of image classification, providing innovative solutions to complex industrial problems. In healthcare, deep learning-based image classification algorithms have made significant progress, enabling the development of highly accurate diagnostic tools. These advances largely improved the diagnostic accuracy and speed up the diagnostic process. The present paper investigates the correlation between the training data scale and classification performance of different models for the skin cancer images.

2.2. Skin Cancer Classification

Classification of skin cancer based on dermoscopic images is very challenging due to time-consuming and subjective manual interpretation by dermatologists. Traditional methods such as biopsy [35], the seven-point checklist [36] and ABCD rule [37] have been used, but they require specialized knowledge and large human efforts [38].

In recent years, computer-aided diagnosis (CAD) [19][39] systems using machine learning algorithms have emerged as a promising approach for automated skin cancer classification [40][41]. These systems are faster and more accurate than human-designed techniques, especially in image segmentation and classification tasks [42][43].

Deep learning algorithms, especially CNNs, have made significant progress in skin cancer classification through lesion segmentation, feature extraction, and high classification accuracy [44][45]. For example, Haenssle et al. [46] analyzed the Google Inception V4 deep learning model. The model outperformed dermatologists in melanoma detection, highlighting the potential of AI-based approaches for skin cancer diagnosis. Pacheco et al. [47] have shown good results in skin lesion classification using deep learning models including GoogleNet, ResNet, VGGNet and MobileNet. Yiming Zhang et al. [48] accomplished the identification of melanoma in skin lesion images on the ISIC2020 dataset using DenseNet [49], with an AUC of 0.925, which is higher than the methods with VGG or ResNet as the backbone.

In addition, transfer learning has been widely used in skin cancer classification to utilize pre-trained models and improve performance [50]. Migration learning allows models to extract features from images that may not be available to the human eye, thus improving sensitivity and specificity compared to dermatologists [51]. In addition, deep learning-based methods have been effective in segmenting skin lesions and detecting and classifying melanomas from dermoscopic images with high accuracy [52][53].

Overall, the integration of deep learning and interpretable transfer learning techniques shows great potential in automatic skin cancer classification, improving diagnostic accuracy and reducing the burden on dermatologists. The present paper investigates the classification performance of different models on the HAM10000 dataset.

2.3. Neural Scaling Laws

Recent advancements in machine learning have witnessed a substantial increase on the accuracies of various machine learning models trained on larger datasets. Recent studies [54][55][56][57][58] have established neural scaling laws [13] in deep neural networks. Notably, the Google team's 2021 study [13] delved into the laws during training. The neural scaling laws refer to a set of empirical rules describing the relationship between the performance of neural networks and their training with different model sizes, dataset scales and computational costs. These rules suggest that as the scale of the training dataset increases, the performance of neural networks generally improves. Additionally, to a certain extent, increasing the number of model parameters and the computational cost significantly, including the number of neurons or layers, would bring clear improvement on model performance. Neural scaling laws play a crucial role in practical applications and advancements across diverse fields, notably in the advancement of large language models such as GPT-1, GPT-2, and GPT-3 developed by OpenAI [59]. These models leverage the principles of neural scaling laws to ascertain the most effective distribution of a fixed computational budget. The present paper validates neural scaling laws on several networks on skin cancer classification.

3. METHODOLOGY

3.1. Deep Image Classification Networks

In this section, four popular types of classification networks were introduced: AlexNet, ResNets, ViT, and MLP-Mixer. These networks have played important roles in advancing computer vision community. The goal is to provide a comprehensive study of the relationship between the performance and dataset scales of the above-mentioned classifiers on the skin cancer classification. The detailed information of computational FLOPs [14] and parameter amounts by different classifiers was provided in Table 1, showcasing their distinct computational characteristics and varied performance on the same dataset. FLOPs [14], or floating-point operations per second, while parameters refer to the number of learnable parameters in a model.

Table 1. The thop [60] package was used in PyTorch to calculate the FLOPs and the number of parameters. The size of input images was 256×256.

Reference	Model	FLOPs	Params
[15]	ResNet18	2G	11M
[15]	ResNet34	5G	22M
[15]	ResNet50	5G	25M
[15]	ResNet101	10G	44M
[17]	MLP-Mixer (B/16)	2G	59M
[15]	ResNet152	15 G	59M
[16]	ViT-Base	17G	102M
[14]	AlexNet	925M	925M

3.1.1. AlexNet

AlexNet [14] is a classic deep learning model widely used for image classification tasks. It was the first deep CNN to achieve significant improvement over traditional models in the ImageNet challenge.

The core structure of AlexNet comprises five convolutional layers and three fully connected layers. The convolutional kernels are of sizes of 11×11, 5×5, 3×3, 3×3 and 3×3, respectively, to increase the network's receptive field. It also uses a large stride and pooling layer to reduce the size of feature maps and improve computational efficiency. In the fully connected layer, AlexNet utilizes Dropout [61] to avoid overfitting and ReLU activation function to enhance the network's nonlinear learning capability. The final fully connected layer employs the Softmax activation function to output the probability vector as the final category prediction. To accommodate with the skin cancer dataset of HAM10000, the output dimension was modified to 7.

AlexNet offers several advantages of simple model structure, high computational efficiency, and good performance on large-scale data. As a result, AlexNet is highly suitable for skin cancer image classification. The architecture of AlexNet is shown in Fig. 1.

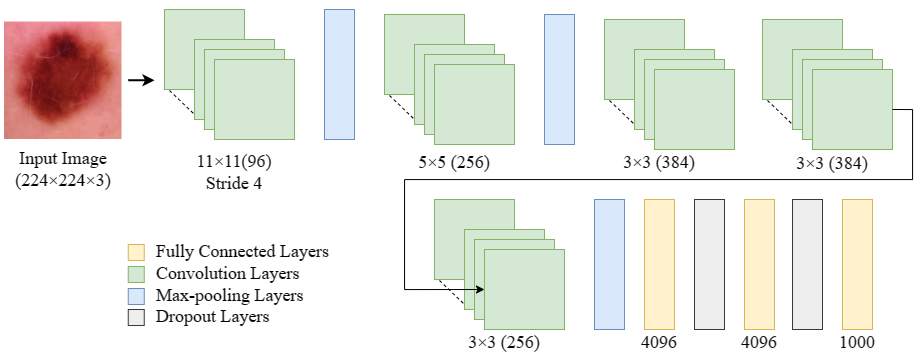

Figure 1. The architecture of AlexNet.

3.1.2. ResNets

ResNets (Residual Network) [15] are a series of deep learning networks designed to mitigate the issues of "gradient vanishing" and "gradient explosion" encountered during the network training [62]. Specifically, ResNet incorporates shortcut connections to directly connect different layers. These connections enable faster propagation of information, more stable gradient flow and the training of deeper networks.

The basic building block of ResNet is the residual block (shown in Fig. 2), which comprises several (two or three) convolutional layers, with batch normalization and ReLU activation (shown in Fig. 3). A shortcut connection is used to add input and output feature, allowing the block to learn residual feature.

ResNet offers several variants with varying depths, such as ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152. Deeper variants have more parameters and can learn more complex representations. By stacking residual blocks, ResNet achieves strong performance on image classification tasks, particularly with large-scale datasets like ImageNet. The architectures of fine-tuned ResNets models in this study are shown in Table 2.

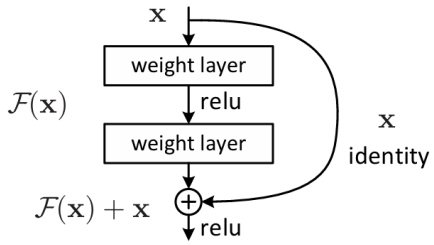

Figure 2. Residual learning: a building block. The image is from [15].

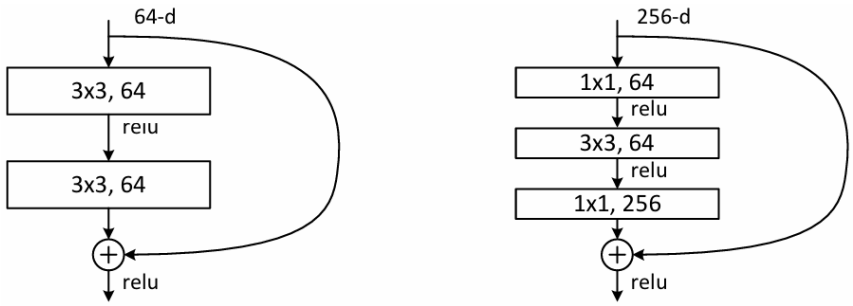

Figure 3. A deeper residual function F for ImageNet. Left: a building block (on 56×56 feature maps) for ResNet34. Right: a "bottleneck" building block for ResNet-50/101/152. The image is from [15].

Table 2. The architecture of ResNets. The table references [15].

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	128×128	7×7, 64, stride 2				
conv2_x	64×64	3×3 max pool, stride 2				
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
conv3_x	32×32	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
conv4_x	16×16	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
conv5_x	8×8	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1×1	average pool, 7-d fc, softmax				

3.1.3. The Vision Transformer (ViT)

The Vision Transformer (ViT) [16] is an image classification network designed from the Transformer [63] architecture. It divides an image into non-overlapping fixed-size patches and converts each patch into a token vector. These vectors are then input to the Transformer encoder, where each attention head can attend to different parts of the image, capturing both local and global dependencies. The core idea of ViT is to leverage Transformer's self-attention mechanism to establish relationships between pixels in an image. ViT contains a stack of transformer layers, each of which has a multi-head self-attention block mechanism and a position-wise feed-forward neural network. The output feature is passed through a fully connected layer for classification. The advantage of ViT compared with traditional CNNs lies in its effective processing of global information in an image with self-attention mechanism. By learning the relationships between image tokens, ViT can achieve promising performance on classification of complex content such as skin cancer images.

The ViT-Base model was utilized in the experiment, which has 12 transformer layers, a hidden size of 768, an MLP size of 3072, and 12 attention heads. The detailed architecture of ViT is shown in Fig. 4.

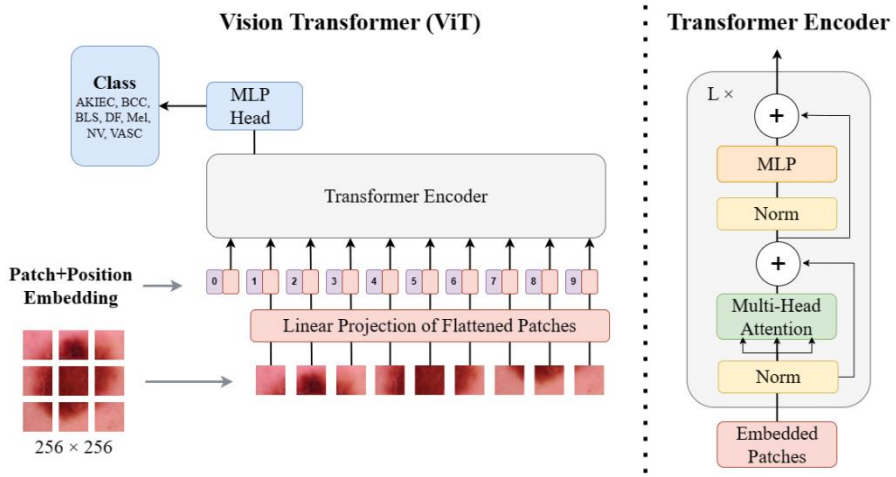

Figure 4. The architecture of ViT. Position embeddings are added to the patch embeddings to retain positional information. The image references [16].

3.1.4. MLP-Mixer

MLP-Mixer [17] is a novel architecture that differs from CNNs and Transformers by avoid using convolutions or self-attention. Instead, it relies on multi-layer perceptron applied sequentially on the spatial and feature channel dimensions. MLP-Mixer consists of two types of MLP layers: channel-mixing MLP and token-mixing MLP, each consisting of two fully connected layers and a non-linear layer, which alternate to facilitate information fusion. It takes non-overlapping image blocks as input, projects them to a desired hidden dimension, and produces a real-valued vector. Unlike most CNNs, the MLP-Mixer has an isotropic design, where each layer receives inputs of the same size. Additionally, unlike ViTs, the MLP-Mixer does not use positional embeddings, as the token mixing MLP can learn to represent positions based on the order of input tokens. It uses a fully connected layers for image classification.

The advantage of MLP-Mixer over previous networks lies in its ability to process global information in images with a stack of complementary spatial-wise and channel-wise feature fusion. This characteristic makes MLP-Mixer well-suited for handling images with complex contents. By modeling the global correlations of the image, MLP-Mixer effectively captures image features and demonstrates strong performance on image classification tasks. Therefore, MLP-Mixer has a wide range of applications in medical image classification [64].

The Base-16 model is a medium-sized MLP-Mixer model, comprising 12 layers, with a patch resolution of 16×16 and a hidden size of 768. Each layer includes an MLP with a channel dimension of 3072 and a hidden layer width of 384. The number of input patches is 196. The Base-16 model was chosen for the experiment as it performs well in general image classification tasks and strikes a balance between model size and performance to a certain extent. The architecture of MLP-Mixer is shown in Fig. 5.

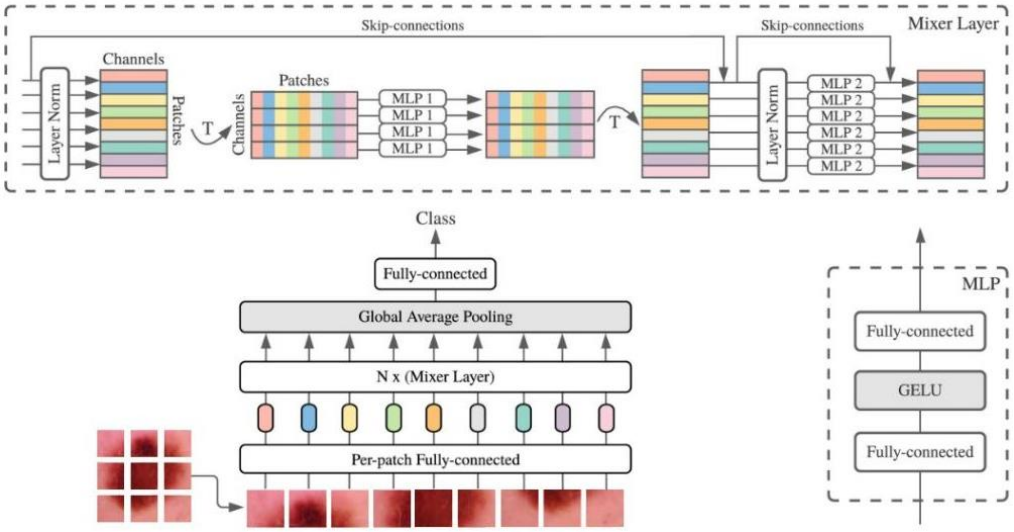

Figure 5. The architecture of MLP-Mixer. The image references [17].

3.1.5. Model Selection and Adaptation

In this study, we choose the extensively employed models like AlexNet, ResNets, ViT-Base, and MLP-Mixer (Base-16 model). These models have demonstrated effective image classification performance. To adapt these models to our experiments, the number of categories is adjusted from 1000 in the original model to 7 in our experiment. Moreover, the input images were resized to 256 × 256 × 3.

3.2. Dataset

The Human Against Machine (HAM10000) dataset [18] is a comprehensive collection of dermatoscopic images including various types of skin lesions. This dataset has become a benchmark in the research community for developing and evaluating skin cancer classification methods. Its diversity and size make it suitable for training and testing deep learning models, enabling researchers to explore new approaches for skin cancer classification [65][66].

The dataset is divided into seven categories based on different types of skin lesions commonly encountered in dermatology: actinic keratosis and intraepithelial carcinoma (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (BLS), dermatofibroma (DF), melanoma (Mel), melanocytic nevi (NV), and vascular lesions (VASC). Fig. 6 shows typical images from the HAM10000 dataset from the seven classes. The dataset specifically avoids ambiguous categorization. Each type of skin lesion has unique characteristics, including appearance, histologic features, and clinical manifestations, making it crucial for accurate diagnosis of skin lesions [18]. The dataset exhibits a significant class imbalance issue, presenting a major challenge for skin cancer classification [65].

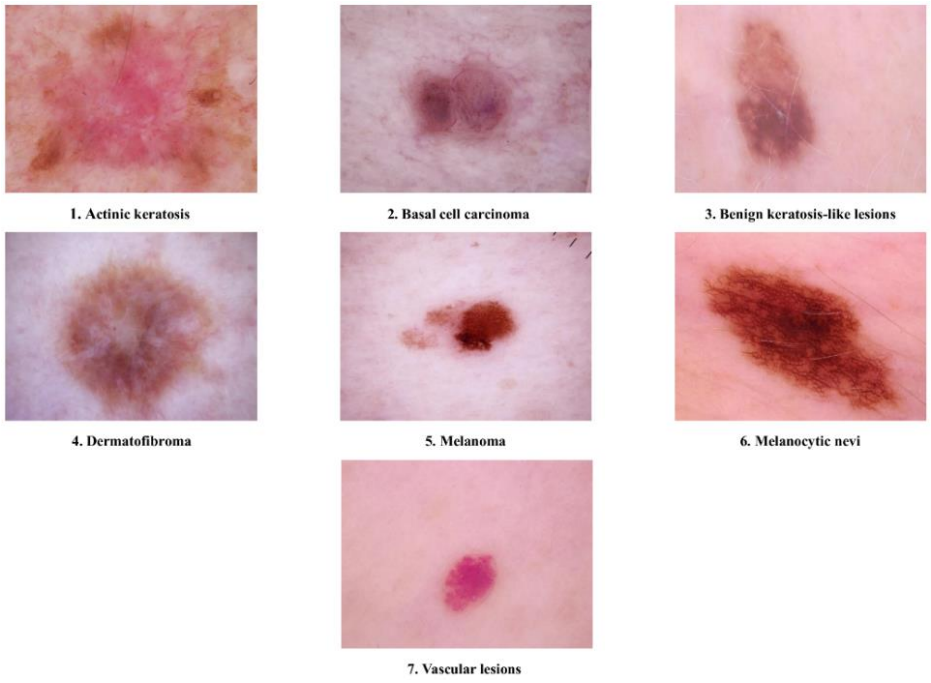

Figure 6. Sample images for the seven skin lesion categories from HAM10000 dataset [18]. Images are of resolution 600 × 450 pixels, saved in JPEG format, and are RGB with three channels.

4. EXPERIMENTS

4.1. Training Details

In this study, eight distinct networks were employed for training: AlexNet, ResNet18, ResNet34, ResNet50, ResNet101, ResNet152, ViT, and MLP-Mixer (Base-16 model). Based on transfer learning, AlexNet and ResNets were pretrained on ImageNet and fine-tuned on the HAM10000 dataset, while the ViT and MLP-Mixer models were trained from scratch with randomly initialized weights. Pretrained models like AlexNet and ResNets could benefit similar tasks like image classification on HAM10000. However, ViT and MLP-Mixer lack pretrained weights trained on ImageNet, necessitating training from scratch.

For the training process, each model was trained for 15 epochs, during which both training and testing losses were monitored. We use the Adam optimizer [32] with a learning rate (LR) [67] of 0.00005 and the Step Learning Rate Scheduler (StepLR) [68] provided by PyTorch with a step size of 5. This scheduler reduces the learning rate by a factor of gamma = 0.5 every 5 epochs to improve training convergence. Dropout with a probability of 0.5 was applied to the fully connected layers of AlexNet and the last linear layer of ResNet models. For ViT and MLP-Mixer models, a dropout rate of 0.1 was applied to the transformer layers as a form of regularization [61].

Data operations included resizing images to 256 × 256 pixels, converting them to tensors. The mean and variance of the HAM10000 dataset were also calculated using PyTorch and then normalizing pixel values using mean = [0.7633, 0.5458, 0.5704] and std = [0.09, 0.1188, 0.1334]. The HAM10000 dataset was divided into seven subsets with sizes of 10015 (100%), 7511 (≈75%), 5008 (≈50%), 2504 (≈25%), 1002 (≈10%), 501 (≈5%), and 100 (≈1%) images, respectively, to evaluate model performance. Datasets were split into training and testing sets with a ratio of 8:2.

The batch size was set to 16, which is a common choice for balancing training efficiency and memory usage [69][70]. The DataLoader used in PyTorch shuffles the training data, introducing randomness and preventing the model from memorizing the order of the data. Four work processes were introduced to load the data in parallel, which can significantly speed up the data loading process, especially for large datasets.

Experiments were conducted using an 11th Gen Intel(R) Core (TM) i7-1165G7 CPU and a GeForce RTX 3060 Ti GPU, operating at a base frequency of 2.80GHz.

4.2. Loss Function

The loss between the predicted outputs and the ground-truth labels is computed using cross-entropy loss function. Assuming the model outputs a vector with N elements, each representing the probability of the sample belonging to the corresponding class, the cross-entropy loss function is calculated as follows:

$$CrossEntropyLoss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Here, y_i is the i-th element of the true label vector, $\hat{y}_i$ is the i-th element of the model's predicted vector, and N is the total number of classes. The objective of the cross-entropy loss function is to minimize the difference between the predicted probability distribution and the true label distribution, thereby improving the model's accuracy in predicting sample classes [71].

5. RESULTS AND DISCUSSION

5.1. Training Fluctuations

As shown in Fig. 7 and Fig. 8, the training process for all models resulted in converged training losses. However, different patterns of test loss fluctuations were observed in different models. Specifically, the AlexNet, ViT, and MLP-Mixer models exhibited a phenomenon in which the test loss initially dropped but then rebounded, suggesting that these models may have been overfitted to the HAM10000 dataset. In contrast, the ResNet series performed best on the HAM10000 dataset, with test losses stabilizing in the range of 0.4 to 0.7 with minimal fluctuations. In addition, the test loss of the ResNet models did not increase significantly over the course of 15 epochs. On the contrary, the ViT model shows the worst convergence with a minimum loss of 0.8, but increases to almost 1.4 at the 15th epoch.

The fluctuations in test loss are relatively small on subsets of 75%, 10%, 5%, and 1%, while larger on the 50% and 25% subsets. These results indicate that while all models achieve convergence in

terms of training loss, the ability to generalize and keep test loss low varies greatly between different models. The ResNet series show better stability and performance on the HAM10000 dataset than the AlexNet, ViT, and MLP-Mixer models.

The fluctuations in loss values observed during training and testing can be attributed to the unbalanced dataset, which can lead to biased models and may affect their performance. None of the models had an accuracy of more than 90%.

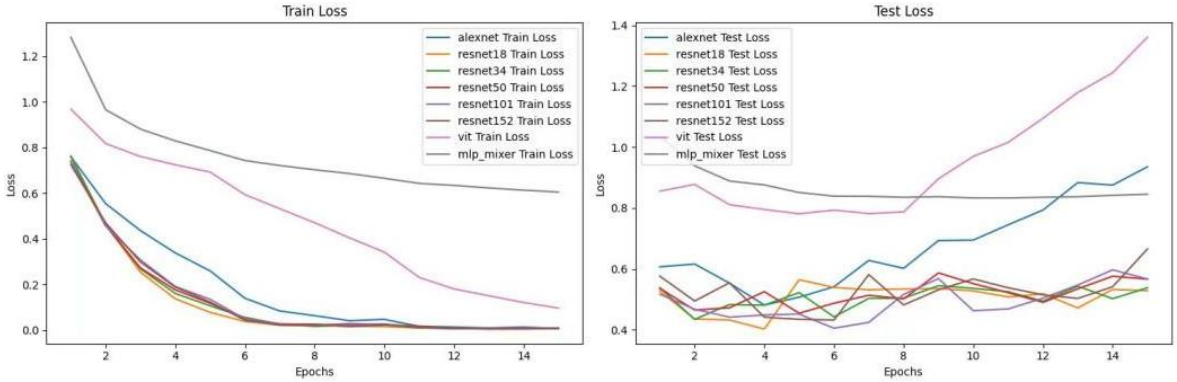

Figure 7. Train loss and test loss for the complete dataset (100%).

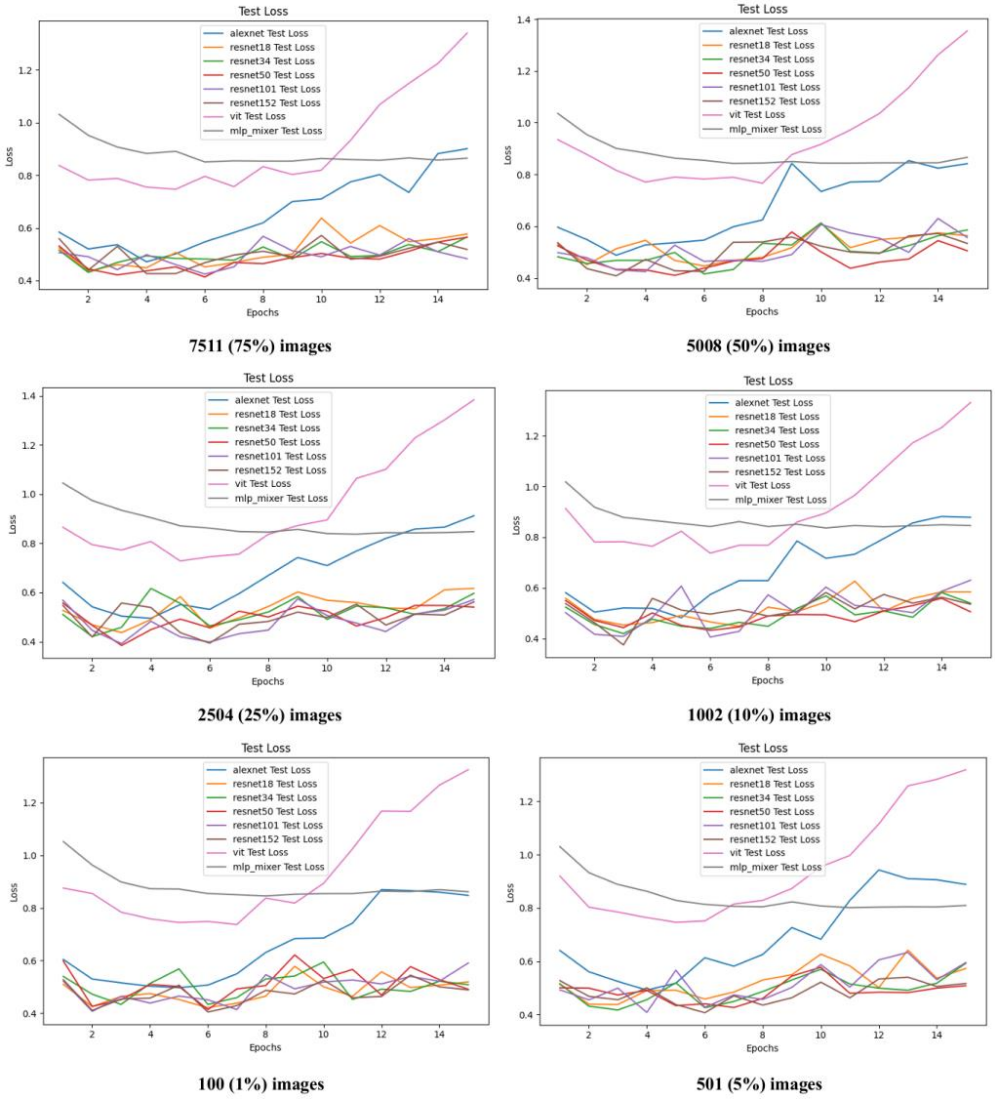

Figure 8. Test loss for 75%, 50%, 25%, 10%, 5%, and 1% subsets.

5.2. Training Analysis

In consideration of using pretrained models versus training from scratch, the results showed that models with pretrained weights (AlexNet and ResNets) generally achieved better performance in terms of both convergence speed and final accuracy compared to models trained from scratch (ViT and MLP-Mixer). Moreover, the training time for models without pretrained weights was longer, as they started with randomly initialized weights and required more epochs to converge.

5.3. Performance Analysis

As shown in Table 1, AlexNet and ResNets generally outperform ViT and MLP-Mixer on the HAM10000 dataset with higher accuracy. Specifically, most of the models achieve the highest performance on the 5% subset, where the average accuracy of all models reaches 85.09%. Notably, ResNet101 on the 100% dataset and ResNet152 on the 5% subset produced the highest accuracy among all cases, reaching 90.07% and 90.02% accuracies, respectively.

This study confirms the effectiveness of neural scaling laws on skin cancer classification. While dataset size is an important factor, our experiments show that more complex models, such as ResNet101 and ResNet152, usually achieve higher accuracies on larger datasets, highlighting the benefits of larger datasets and model complexity. These results suggest that the relationship between dataset size and model performance is multifaceted. While larger datasets provide more information, the effectiveness of the model also depends on its complexity, the distribution of the dataset, and the efficiency of model training [13]. Therefore, factors such as model architecture and training strategy should be carefully considered in addition to dataset size when developing skin cancer classification methods. Detailed information on FLOPs and the number of parameters for all models is shown in Table 3.

Table 3. Accuracy with different subset sizes.

Models	Accuracy (%)							
	10015 (100%)	7511 (75%)	5008 (50%)	2504 (25%)	1002 (10%)	501 (5%)	100 (1%)	average
ResNet18	89.23	88.28	88.43	88.68	88.38	88.28	88.58	88.55
ResNet34	88.33	88.63	89.18	89.23	89.13	89.38	89.48	89.05
ResNet50	89.03	89.48	89.98	89.78	88.88	89.98	89.28	89.49
ResNet101	90.07	89.48	89.38	88.68	89.08	89.53	89.48	89.39
MLP	70.22	70.12	70.87	71.32	70.57	71.97	70.72	70.83
ResNet152	89.58	89.48	89.48	89.58	89.58	90.02	88.83	89.51
ViT	74.46	74.51	73.97	74.51	74.46	74.91	74.01	74.40
AlexNet	85.79	86.43	86.43	85.89	86.68	86.68	85.64	86.22
average	84.59	84.55	84.72	84.71	84.59	85.09	84.50	84.68

Longitudinal comparisons of the ResNets models (from ResNet18 to ResNet152) reveal a general trend of performance improvement with increasing model depth. Deep networks can learn more complex and abstract features, enhancing their classification capabilities. However, our experiments also demonstrated some performance fluctuations, such as a slight decrease in test accuracy between ResNet50 and ResNet101, which could be attributed to overfitting due to increased model complexity [72]. In a horizontal comparison, CNN is good at capturing local features of an image, while attention-based models are better suited for tasks involving global features [15][61][73].

5.4. Limitations of Our Work

Firstly, our work is influenced by the constraints of our dataset, limiting our ability to perform predictive classification on a larger dataset. Consequently, we are unable to analyze the performance of models under conditions involving larger datasets.

Secondly, these models did not exhibit a clear trend on the dataset. This could be attributed to the inherent robustness of these models.

Thirdly, the test set size was not fixed, resulting in fluctuating evaluation metrics. The reason is that, the number of training samples should match the number of test samples in a corresponding proportion, as this better reflects real-world scenarios in hospitals.

6. CONCLUSION

This study investigated the performance of eight deep neural networks in skin cancer image classification. The results on the HAM10000 dataset demonstrated that, the complex interplay between data size, model complexity, and computational costs, validated the effectiveness of neural scaling laws in this problem. The models were trained on datasets ranging from 1% to 100%, with the highest average accuracy of 85.09% when the dataset size was 5% (1002 images). Notably, ResNet101 trained with 100% dataset and ResNet152 trained with 5% dataset obtained the highest accuracy of 90.07% and 90.02% respectively. Our experimental results suggest that the dataset size and model complexity both significantly affect performance. For example, ResNet101 and ResNet152 achieved higher accuracy on larger datasets. In contrast, AlexNet, ViT, and MLP-Mixer performed best on smaller datasets. By understanding the impact of dataset size and model complexity, researchers can develop more efficient and accurate skin cancer classification methods. Overall, this study promotes advances in skin cancer classification methods that have the potential to improve healthcare outcomes and reduce the burden on healthcare professionals.

REFERENCES

[1] Li W., Raj A.N.J., Tjahjadi T., Zhuang Z. Digital hair removal by deep learning for skin lesion segmentation. Pattern Recognit. 2021; 117:107994. doi: 10.1016/j.patcog. 2021.107994.

[2] Karimkhani C., Green A.C., Nijsten T., Weinstock M.A., Dellavalle R.P., Naghavi M., Fitzmaurice C. The global burden of melanoma: results from the Global Burden of Disease Study 2015. Br J Dermatol. 2017 Jul; 177(1):134-140. doi: 10.1111/bjd.15510.

[3] Oliveira R.B., Filho M.E., Ma Z., Papa J.P., Pereira A.S., Tavares J.M.R. Computational methods for the image segmentation of pigmented skin lesions: A review. Comput. Methods Programs Biomed. 2016; 131:127–141. doi: 10.1016/j.cmpb.2016.03.032.

[4] M. Elbaum et al. Automatic differentiation of melanoma from melanocytic nevi with multispectral digital dermoscopy: A feasibility study. J Am Acad Dermatol (2001).

[5] Reshma G., Al-Atroshi C., Nassa V.K., Geetha B., Sunitha G., Galety M.G., Neelakandan S. Deep Learning-Based Skin Lesion Diagnosis Model Using Dermoscopic Images. Intell. Autom. Soft Comput. 2022; 31:621–634. doi: 10.32604/iasc.2022.019117.

[6] Brinker T.J., Hekler A., Enk A.H., Berking C., Haferkamp S., Hauschild A., Weichenthal M., Klode J., Schadendorf D., Holland-Letz T., von Kalle C., Fröhling S., Schilling B., Utikal J.S. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer. 2019 Sep; 119:11-17. doi: 10.1016/j.ejca.2019.05.023.

[7] Pomponiu V, Nejati H, Cheung NM. Deepmole: Deep neural networks for skin mole lesion classification. In: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP).

[8] Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr SMR, Jafari MH, Ward K, et al. Melanoma detection by analysis of clinical images using convolutional neural network. In: Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

[9] I. Giotis, N. Molders, S. Land, M. Biehl, M.F. Jonkman and N. Petkov: "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images", Expert Systems with Applications, 42 (2015), 6578-6585

[10] Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2024). Demystification and Actualisation of Data Saturation in Qualitative Research Through Thematic Analysis. International Journal of Qualitative Methods, 23. https://doi.org/10.1177/16094069241229777

[11] O'Reilly, M., & Parker, N. (2013). 'Unsatisfactory Saturation': a critical exploration of the notion of saturated sample sizes in qualitative research. Qualitative Research, 13(2), 190-197. https://doi.org/10.1177/1468794112446106

[12] Esteva, A., Robicquet, A., Ramsundar, B. et al. A guide to deep learning in healthcare. Nat Med 25, 24–29 (2019). https://doi.org/10.1038/s41591-018-0316-z

[13] Bahri, Y., Dyer, E., Kaplan, J., Lee, J., Sharma, U. Explaining Neural Scaling Laws. ArXiv. 2021. DOI: 10.48550/arXiv.2102.06701.

[14] Krizhevsky, A., Sutskever, I., Hinton, G. ImageNet classification with deep convolutional neural networks. Commun ACM. 2012; 60:84–90.

[15] He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016;770-778.

[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR). 2020.

[17] Tolstikhin, IO., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., et al. Mlp-mixer: An all-mlp architecture for vision. In: Advances in neural information processing systems. 2021; 34:24261-24272.

[18] Tschandl, P., Rosendahl, C., Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data. 2018; 5:180161. doi: 10.1038/sdata.2018.161.

[19] Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep Learning in Medical Image Analysis. Adv Exp Med Biol. 2020; 1213:3-21. doi: 10.1007/978-3-030-33128-3_1. PMID: 32030660; PMCID: PMC7442218.

[20] Wu, M., Zhou, J., Peng, Y., Wang, S., Zhang, Y. (2024). Deep Learning for Image Classification: A Review. In: Su, R., Zhang, YD., Frangi, A.F. (eds) Proceedings of 2023 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2023). MICAD 2023. Lecture Notes in Electrical Engineering, vol 1166. Springer, Singapore. https://doi.org/10.1007/978-981-97-1335-6_31

[21] Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN COMPUT. SCI. 2, 420 (2021). https://doi.org/10.1007/s42979-021-00815-1

[22] Amos, Brandon, Bartosz Ludwiczuk and Mahadev Satyanarayanan. "OpenFace: A general-purpose face recognition library with mobile applications." (2016). https://api.semanticscholar.org/CorpusID:16506546

[23] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823). DOI: 10.1109/CVPR.2015.7298682

[24] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4690-4699). DOI: 10.1109/TPAMI.2021.3087709

[25] Priit, M., Chern, G. (2022). Transfer Learning in Satellite Imagery Classification: A Comparative Study of Custom CNN and Pre-trained Models. arXiv preprint arXiv:2010.06497.

[26] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

[27] Quinlan, J.R. Induction of decision trees. Mach Learn 1, 81–106 (1986). https://doi.org/10.1007/BF00116251

[28] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[29] Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9

[30] Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. Front Public Health. 2023 Nov 7; 11:1273253. doi: 10.3389/fpubh.2023.1273253. PMID: 38026291; PMCID: PMC10662291.

[31] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2014). Going Deeper with Convolutions. arXiv:1409.4842. https://doi.org/10.48550/arXiv.1409.4842

[32] Kingma, DP., Ba, J. Adam: A method for stochastic optimization. In: Proceedings of the 2015 International Conference on Learning Representations (ICLR); San Diego, CA, USA. 2015. (pp. 1–15).

[33] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). Mixup: Beyond Empirical Risk Minimization. arXiv:1710.09412. https://doi.org/10.48550/arXiv.1710.09412.

[34] Farahani, A., Pourshojae, B., Rasheed, K., & Arabnia, H. R. (2021). A Concise Review of Transfer Learning. arXiv:2104.02144. https://doi.org/10.48550/arXiv.2104.02144

[35] Kilic A, Kilic A, Kivanc AE, Sisik A. Biopsy Techniques for Skin Disease and Skin Cancer: A New Approach. J Cutan Aesthet Surg. 2020 Jul-Sep;13(3):251-254. doi: 10.4103/JCAS.JCAS_173_19. PMID: 33209007; PMCID: PMC7646420.

[36] Argenziano G., Fabbrocini G., Carli P., De Giorgi V., Sammarco E., Delfino M. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Arch Dermatol. 1998 Dec;134(12):1563–70. doi: 10.1001/archderm.134.12.1563.

[37] Senan E.M., Jadhav M.E. Analysis of dermoscopy images by using ABCD rule for early detection of skin cancer. Global Transitions Proc. 2021; 2:1–7. doi: 10.1016/j.gltp.2021.01.001.

[38] Barata C., Celebi M.E., Marques J.S. A survey of feature extraction in dermoscopy image analysis of skin cancer. IEEE J Biomed Health Inform. 2018; 23:1096–109. doi: 10.1109/JBHI.2018.2845939.

[39] Ningrum, DNA., Yuan, SP., Kung, WM., Wu, CC., Tzeng, IS., Huang, CY., et al. Deep learning classifier with patient's metadata of dermoscopic images in malignant melanoma detection. J Multidiscip Healthc. 2021; 14:877–885. doi: 10.2147/JMDH.S306284.

[40] Masood, A., Ali Al-Jumaily, A. Computer-aided diagnostic support system for skin cancer: a review of techniques and algorithms. Int J Biomed Imaging. 2020; 2020:323268. doi: 10.1155/2013/323268.

[41] Mobiny, A., Singh, A., Van Nguyen, H. Risk-aware machine learning classifier for skin lesion diagnosis. J Clin Med. 2019; 8:1241. doi: 10.3390/jcm8081241.

[42] Kassem, MA., Hosny, KM., Fouad, MM. Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning. IEEE Access. 2020; 8:114822–114832. doi: 10.1109/ACCESS.2020.3003890.

[43] Szaleniec, J., Szaleniec, M., Stręk, P., Boroń, A., Jabłońska, K., Gawlik, J., et al. Outcome prediction in endoscopic surgery for chronic rhinosinusitis–a multidimensional model. Adv Med Sci. 2014; 59:13–18. doi: 10.1016/j.advms.2013.06.003.

[44] Khasawneh, N., Fraiwan, M., Fraiwan, L., Khassawneh, B., Ibnian, A. Detection of COVID-19 from Chest X-ray Images Using Deep Convolutional Neural Networks. Sensors. 2021; 21:5940. doi: 10.3390/s21175940.

[45] Shorten, C., Khoshgoftaar, TM. A survey on Image Data Augmentation for Deep Learning. J Big Data. 2019;6. doi: 10.1186/s40537-019-0197-0.

[46] Kadampur, MA., Al Riyaee, S. Skin cancer detection: Applying a deep learning-based model-driven architecture in the cloud for classifying dermal cell images. Inform Med Unlocked. 2020; 18:100282. doi: 10.1016/j.imu.2019.100282.

[47] Li, Lingyun, Xu Wang, Weijian Hu, Neal Naixue Xiong, Yongxing Du and Baoshan Li. "Deep Learning in Skin Disease Image Recognition: A Review." IEEE Access 8 (2020): 208264-208280.

[48] Y. Zhang and C. Wang, "SIIM-ISIC Melanoma Classification With DenseNet," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 2021, pp. 14-17, doi: 10.1109/ICBAIE52039.2021.9389983.

[49] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. arXiv preprint arXiv:1608.06993 [cs.CV]. Retrieved from https://doi.org/10.48550/arXiv.1608.06993

[50] Fraiwan, M., Faouri, E. On the Automatic Detection and Classification of Skin Cancer Using Deep Transfer Learning. Sensors (Basel). 2022; 22:4963. doi: 10.3390/s22134963.

[51] Efimenko, M., Ignatev, A., Koshechkin, K. Review of medical image recognition technologies to detect melanomas using neural networks. BMC Bioinformatics. 2020;21(Suppl 11):270. doi: 10.1186/s12859-020-03615-1.

[52] Haenssle, HA., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., et al. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018; 29:1836–1842. doi: 10.1093/annonc/mdy166.

[53] Jinnai, S., Yamazaki, N., Hirano, Y., Sugawara, Y., Ohe, Y., Hamamoto, R. The Development of a Skin Cancer Classification System for Pigmented Skin Lesions Using Deep Learning. Biomolecules. 2020; 10:1123. doi: 10.3390/biom10081123.

[54] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., & Zhou, Y. (2017). Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

[55] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

[56] Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., & Shavit, N. (2020). A constructive prediction of the generalization error across scales. In International Conference on Learning Representations.

[57] Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., & McCandlish, S. (2020). Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701.

[58] Rosenfeld, J. S., Frankle, J., Carbin, M., & Shavit, N. (2020). On the predictability of pruning across scales. arXiv preprint arXiv:2006.10621.

[59] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Radford, A. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

[60] GitHub - Lyken17/pytorch-OpCounter: Count the MACs / FLOPs of your PyTorch model.

[61] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014; 15:1929-1958.

[62] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. Gradient-based learning applied to document recognition. In: Proceedings of the IEEE. 1998.

[63] Vaswani, Ashish et al. "Attention is All you Need." Neural Information Processing Systems (2017). arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762.

[64] Liu Y, Xing W, Zhao M, Lin M. A new classification method for diagnosing COVID-19 pneumonia based on joint CNN features of chest X-ray images and parallel pyramid MLP-mixer module. Neural Comput Appl. 2023 Apr 28:1-13. doi: 10.1007/s00521-023-08604-y. Epub ahead of print. PMID: 37362575; PMCID: PMC10147369.

[65] Yao, P., Shen, S., Xu, M., Liu, P., Zhang, F., Xing, J., et al. Single Model Deep Learning on Imbalanced Small Datasets for Skin Lesion Classification. arXiv. 2021. doi: 10.1109/TMI.2021.3136682.2102.01284.

[66] Shetty, B., Fernandes, R., Rodrigues, A.P. et al. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. Sci Rep 12, 18134 (2022). https://doi.org/10.1038/s41598-022-22644-9

[67] Hassan, E., Shams, MY., Hikal, NA., Elmougy, S. The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study. Multimed Tools Appl. 2023; 82:16591-16633. doi: 10.1007/s11042-022-13820-0.

[68] PyTorch Contributors. (2023). PyTorch: StepLR. https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.StepLR.html

[69] You, Y., et al. Large Batch Training of Convolutional Networks. ArXiv. 2017.

[70] Shen, L., et al. On Efficient Training of Large-Scale Deep Learning Models: A Literature Review. ArXiv. 2023.

[71] Mao, A., Mohri, M., & Zhong, Y. (2023). Cross-Entropy Loss Functions: Theoretical Analysis and Applications. Retrieved from https://doi.org/10.48550/arXiv.2304.07288

[72] Rezaeezade, A., Perin, G., Picek, S. To Overfit, or Not to Overfit: Improving the Performance of Deep Learning-Based SCA. In: Batina, L., Daemen, J. (eds) Progress in Cryptology - AFRICACRYPT 2022. AFRICACRYPT 2022. Lecture Notes in Computer Science, vol 13503. Springer, Cham. 2022. https://doi.org/10.1007/978-3-031-17433-9_17.

[73] Hang, R., Li, Z., Liu, Q., Ghamisi, P., Bhattacharyya, SS. Hyperspectral Image Classification With Attention-Aided CNNs. IEEE Trans Geosci Remote Sens. 2021; 59:2281-2293. doi: 10.1109/TGRS.2020.3007921. (pp. 1-4).