CT Image Classification for Lung Cancer

Manhui Zhu
manhuiz@ad.unc.edu
Zeqi Zhou
zeqi@ad.unc.edu
Dehan Cui
dehan@live.unc.edu
Boyu Liu
boyuliu01@unc.edu
University of North Carolina at Chapel Hill

Abstract

Lung cancer is one of the diseases with the highest mortality rate. The most promising way to enhance the survival rate of patients is early detection of cancer. The accurate diagnosis of pathological tumor type of lung cancer is significant for early detection and corresponding treatment. An important and most common criteria for determining tumor type is via computed tomography (CT) data. Traditionally, only radiologists are able to read and interpret the CT scan and provide a diagnosis. However, a qualified radiologist is usually required to complete over 10-years training and schooling. In this paper, we use lung cancer dataset from the *Iraq-Oncology teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD)* for image classification on 3 different convoluted neural networks (CNNs) to determine the tumor type. The methods we used are **VGG-16**, **ResNet-50 and 101**, and **Vision Transformer (ViT)**. The highest accuracy we got is 100.00% from ViT. To improve the model performance, we used the **curriculum learning**, a training technique that trains machine learning models from easier data to harder data, to pre-organized our dataset and then feed the model. We try curriculum learning on ResNet-50 and 101 models, and the test accuracy increased about at least 5%. Here is a link of the code: `https://github.com/BoyuL129/STOR-566`

1 Introduction

Lung cancer accounts for almost 25% of all cancer deaths [1]. Over the past five years, the survival rate is 25% nationally. And close to 237,000 people will be diagnosed with lung cancer in 2022. Lung cancer has one of the lowest five-year survival rates because cases are often diagnosed at a later stage, when the disease is less likely to be curable. The five-year survival rate is higher (61%) if the patient is diagnosed at an early stage. Unfortunately, only 26% of cases are found at an early stage, 44% of cases are not caught until a late stage when the survival rate is at 7% [2]. If lung cancer is found at an earlier stage, it is more likely to be treated successfully.

Computed tomography (CT) scan is commonly used to help the doctor to give a diagnosis. It uses x-ray to make detailed, cross-sectional images of a patient's body, which is more likely to show lung tumors than routine chest x-ray. Instead of taking 1 or 2 pictures like regular x-ray, the CT scan takes many pictures from different angles and then the computer combines them to show a slice of the part of the body being studied. The CT scan can show the size, shape, and position of lung tumors and can help find enlarged lymph nodes that might contain cancer that has spread.

Traditionally, radiologists, who are specialized in diagnosing and treating injuries and diseases using medical imaging procedures, interpret patients' physical conditions according to the CT scan.

Preprint. Under review.

However, a qualified radiologist needs to complete over 10 years of training, including medical school, a four-year residency, and most often, an additional one- or two-year fellowship of very specialized training. Therefore, radiologists are among the most in-demand physician specialists in the U.S. Fortunately, radiology is a data-centric field, which provides the extraction and quantitative features to quantify the solid tumor radiographic phenotype. The mass data of radiology provides the foundation for model training. In recent years, with large amounts of data and hardware advances in high resolution image acquisition equipment, various novel artificial intelligence (AI) algorithms and emerging machine learning and deep learning models are widely used in the medical diagnosis field. Convolutional neural networks (CNNs) have promising performance in their ability to classify images and detect objects and features from images. The deep learning methods can automatically learn and identify complex patterns in replace of the often-subjective assessment of images by trained clinicians and radiologists. In this work, we applied three popular image classification models, VGG-16, ResNet-50 and 101, and Vision Transformer, on our dataset to classified CT scans for lung cancers. Then we use curriculum learning to see whether it can effectively improve our model performance. We arrange our training dataset from easy to difficult according to their loss and feed the model at a specific pace.

2 Related Works

2.1 Cancer Detection

Over the last few years, experts and scholars developed various techniques and models for problems such as clinical detection, classification, and staging of tumors. One hotspot among all research topics is using convolutional neural networks (CNN) to detect and classify lung cancer. CNN has made significant progress in image processing, image recognition, and other fields. There are various architectures made by researchers and scholars in a wide range of computer vision tasks, and some popular models, such as VGG16, ResNet, ViT, and Inception, are used for cancer detection. However, directly using these models on CT scans only gives an ideal result. Therefore, some researchers try to improve existing models to make it more suitable for cancer detection. For example, Professor Lei Geng and his team developed a lung parenchymal segmentation algorithm based on the combination of VGG-16 and dilated convolution [3]. And the improved algorithm shows significant improvement on accuracy compared to original methods. Since the previous work has proven that specialized improvement on existing methods could potentially raise the accuracy, in this paper, we will combine existing image classification models with curriculum learning algorithms to improve its performance.

2.2 Curriculum Learning

Over the past few years, various curriculum learning algorithms have been proposed to enhance the accuracy of image classification. In *On the Power of Curriculum Learning in Training Deep Networks*, Guy Hacohen and Daphna Weinshall design a curriculum that involves a non-uniform sampling of mini-batches to train networks. [4] They decompose their curriculum learning into two tasks: **(i)** Define scoring functions to sort each sample in the data by its difficulty, and **(ii)** develop pacing functions to determine the pace by which data is presented to the network. The algorithms with curriculum learning display noticeable improvement (in terms of final accuracy) in image classification from vanilla algorithms. While their curriculum shows remarkable results, the dataset Guy and Daphna use are CIFAR-10 and CIFAR-100 instead of cancer images.

3 Proposed Methods

3.1 VGG-16

VGG-16 is a convolutional neural network that is 16 layers deep. It was first presented in the publication "Very Deep Convolutional Networks for Large-Scale Image Recognition" by K. Simonyan and A. Zisserman of the University of Oxford. In ImageNet, a dataset of over 14 million pictures classified into 1000 classes, the model achieves 92.7% top-5 test accuracy. It was one of the well-known models submitted to the ILSVRC-2014. It outperforms AlexNet by substituting huge kernel-sized filters (11 and 5, respectively, in the first and second convolutional layers) with numerous

33 kernel-sized filters one after the other. VGG16 had been training for weeks on NVIDIA Titan Black GPUs.

The input to the cov1 layer is a 224 × 224 RGB picture. The picture is processed through a stack of convolutional (conv.) layers, with the filters set to catch the notions of left/right, up/down, and center with a very tiny receptive field: 33 (the smallest size to capture the notions of left/right, up/down, and center). It also employs 11 convolution filters in one of the configurations, which may be thought of as a linear modification of the input channels (followed by non-linearity). The convolution stride is fixed at 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is kept after convolution, i.e. the padding is 1-pixel for 33 conv. Layers. Five max-pooling layers follow some of the conv. layers and do spatial pooling (not all the conv. layers are followed by max-pooling). Max-pooling is done with stride 2 across a 22 pixel frame. Following a stack of convolutional layers (with varying depths in various designs), three Fully-Connected (FC) layers are added: the first two have 4096 channels each, the third conducts 1000-way ILSVRC classification and hence comprises 1000 channels (one for each class). The soft-max layer is the last layer. In all networks, the configuration of the completely linked layers is the same. The rectification (ReLU) non-linearity is present in all buried layers. It is also worth noting that none of the networks (save one) use Local Response Normalisation (LRN), which does not enhance performance on the ILSVRC dataset but increases memory usage and computation time. [5]

3.2 ResNet

Deep Residual network, shortened as Resnet, is designed to train deep networks. Inspired by excellent performance of "very deep" networks on the ImageNet dataset, researchers proposed a question: "Is learning better networks as easy as stacking more layers?" After experimenting with 20-layer networks and 56-layer networks on the CIFAR-10 dataset, they found the 56-layer networks had higher training error and testing error. Eventually the degradation problem associated with deep networks arose: with the network depth increasing, accuracy gets saturated and then degrades rapidly.

Microsoft proposed a deep residual learning architecture to address this issue. Instead of assuming that every few stacked layers will match a desired underlying mapping directly, they expressly allow these layers to fit a residual mapping. Feedforward neural networks with shortcut connections can implement the formulation of

$$F(x) + x \tag{1}$$

Shortcut connections are those that skip one or more of the levels depicted in Figure 1. The shortcut connections execute identity mapping, and their outputs are added to the stacked layers' outputs. Many issues may be handled by employing the residual network, including: ResNets are simple to optimize, however "basic" networks (which just stack layers) exhibit increasing training error as depth increases; ResNets can easily gain accuracy from greatly increased depth, producing results which are better than previous networks. [6]

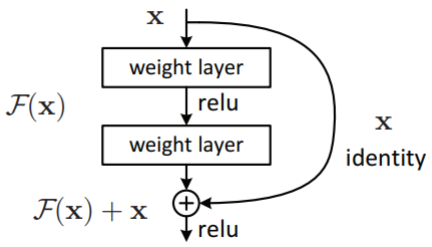

Figure 1: Residual Learning: a learning block [6].

Based on a 34-layer plain network, a shortcut link is placed to transform the network into its corresponding residual version. When the input and output have the identical dimensions, the identity shorthand

$$F(x, Wi) + x \tag{2}$$

can be utilized directly. When the dimensions rise (as seen by the dotted lines in Fig. 2), it considers two options: the shortcut performs identity mapping, with extra zero entries padded for increasing

dimensions. This option introduces no additional parameter.Or, to match dimensions, the projection shortcut in the above equation is applied.

The image is resized with its shorter side randomly sampled in [256, 480] for scale augmentation. A 224×224 crop is randomly sampled from an image or its horizontal flip, with the per-pixel mean subtracted. The learning rate starts from 0.1 and is divided by 10 when the error plateaus and the model is trained for up to 60×10000 iterations by using a weight decay of 0.0001 and a momentum of 0.9.

3.3 Vision Transformer (ViT)

Transformers are initially used for machine translation which replace the recurrence and convolutions entirely with self-attention mechanisms and achieve outstanding performance. Later, transformers became the dominant models for various natural language processing (NLP) tasks. Motivated by their success on the NLP tasks, recent researchers attempted to combine the self-attention mechanism into CNNs for computer vision tasks. Those achievements also stimulate interests of the community in building purely transformer-based models (without convolutions and inductive bias) for vision tasks. Vision Transformer is the first such example of a transformer-based method to match or even surpass CNNs for image classification. Many variants of vision transformers have also been recently proposed that use distillation for data-efficient training of Vision Transformer, pyramid structure like CNNs, or self-attention to improve the efficiency via learning an abstract representation instead of performing all-to-all self-attention. [7]

3.4 Curriculum Learning

Curriculum learning is a technique used to boost the final testing accuracy of our neural networks. The design of curriculum learning mimics the design of human curriculum. For human curriculum, teachers have to determine the order by which the materials are presented to the students and how fast they teach. The two components correspond to the two key aspects of curriculum learning: a scoring function to measure the difficulty of the training examples and a pacing function to determine the pace by which the training examples are presented to the networks.

Normally, the design of scoring functions in curriculum learning requires transfer learning. That is, The design involves models trained in other datasets to generate predictions for our own dataset. The predictions and the true labels of the training set will be combined to create a measure for the difficulty of the training examples. The training examples are then sorted based on the difficulty scores. The lower the score, the easier a particular training example is. Conventionally, the training examples are sorted in an ascending order from the easiest to the hardest.

In classifying lung cancer images, the scoring function is the Sparse Categorical Cross Entropy. We first used the pre-trained MobileNet model on ImageNet to generate Sparse Categorical Cross Entropy on the training set. We calculated the losses with the true labels of the training examples and the predictions generated by the MobileNet model. We then regarded the losses as the difficulty scores for the training examples. The higher the loss for a particular example, the more difficult the example is. Then, we sorted the training examples from the easiest to the hardest based on the difficulty scores. The newly sorted training set is the dataset we use to train our models. For the pacing function, we decided to use a uniform sampling method. We sampled mini-batches with fixed size to feed to the networks.

4 Experiment

We evaluate our method on the *Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD)* lung cancer dataset that consists of 110 patients' CT scans. The 110 cases vary in gender, age, educational attainment, area of residence, and living status to eliminate unwanted skewness as much as possible. Moreover, all identities were waived before utilizing this project. Our group also manually removed labels and timestamps to exclude confounding factors that may affect the performance of the analysis. The patients are grouped into three classes: 55 normal cases, 15 benign cases, 40 malignant cases. For each case, about 10 representing CT scan slices of human chest with different sides and angles are selected from the CT image sequence. In total, it provides

1097 CT scans. In our research, we split the dataset into a training set (70%), a testing set (20%), and a validation set(10%).

Each model is trained on the 768 training images within 50 epochs, and evaluated on the 110 validation images. We also evaluate performance of each model on the 219 test images to obtain final accuracy. Table 1 is a summary of our results. And the Figure 2-5 are accuracy and loss for the models we used.

After the group presentation, and the discussion with our instructor, we think our project should include some state-of-art algorithm with no limitation on input image size to see how curriculum learning strategy would affect their performance. Thus, we replaced the inception-v3 algorithm with Vision Transformer. We applied curriculum learning on ResNet-50, ResNet-101, ViT, and VGG-16. We first used the pre-trained MobileNet on ImageNet to calculate the Sparse Categorical Cross Entropy of the training examples. Then, based on the loss of each example, we sorted the examples from the easiest to the hardest, fed the sorted training dataset to ResNet50, ResNet101, and VGG16, and trained 50 epochs on each model. The results are shown below.

We see a boost in final testing accuracies in the models. ResNet50's accuracy increased from 82.49% to 92.75%, ResNet101's accuracy increased from 82.69% to 87.39% and VGG16's accuracy increased from 87.78% to 92.33%. The losses with curriculum learning converged to 0 more quickly than the losses without curriculum learning. The training accuracies with curriculum learning also converged to higher values than the accuracies without curriculum learning. The Figure 6-7 are loss and accuracy for ResNet-50 and ResNet-101 after applying the curriculum learning.

As mentioned above, we added Vision Transformer as our model. However, since our dataset is not originally selected for Vision Transformer, its training accuracy and loss converge so fast that it achieves 0.9773 accuracy and 0.0479 loss after first epoch, and after second epoch, its accuracy remains at 1.000 and its loss decrease to 0.0000. Thus, applying curriculum learning to Vision Transformer did not make any difference on its accuracy and loss.

Table 1: Test Accuracy

Models	Test Accuracy
VGG-16	87.78%
ResNet-50	82.49%
ResNet-101	82.69%
ViT	100.00%

5 Discussion

For future improvements, we think of three aspects to improve: our models' adaptability, accuracy, and multi-function. We hope our models can be enhanced to be capable of processing various types of medical images other than CT scans. The models can also be improved in terms of accuracy by more delicately designed curriculum learning algorithms, such as using exponential pacing functions. Hopefully, the models can also be applied to classify cancer based on stages and types of tumors and locate it.

Even though curriculum learning can improve the accuracy of some popular CNN models, it is not helpful to state-of-the-art methods, such as ViT. In our experiment, the ViT's loss and train accuracy converge in the first three epochs. It converges so rapidly that it is meaningless to improve it with curriculum learning. However, it does not necessarily mean curriculum learning cannot combine with ViT. Our experiment is limited by our computing power and time, but when solving real world problems, ViT requires a large amount of datasets to achieve top performance. And curriculum learning has the potential to significantly increase its efficiency while training on large datasets. Thus, with more time and better computers, our project can be further extended.

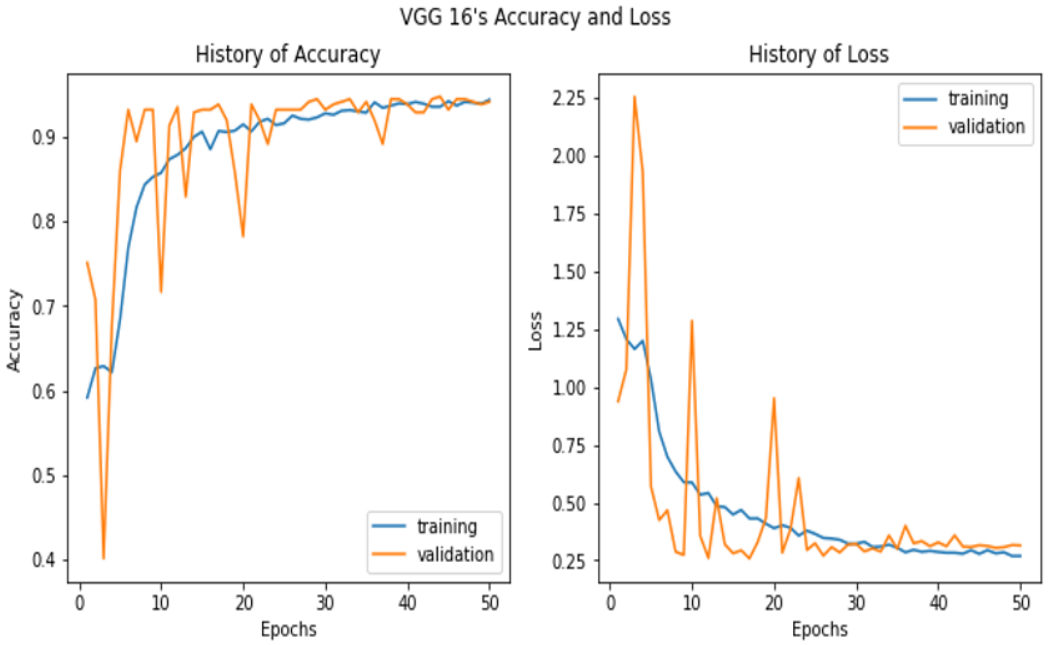

Figure 2: VGG-16's Accuracy and Loss.

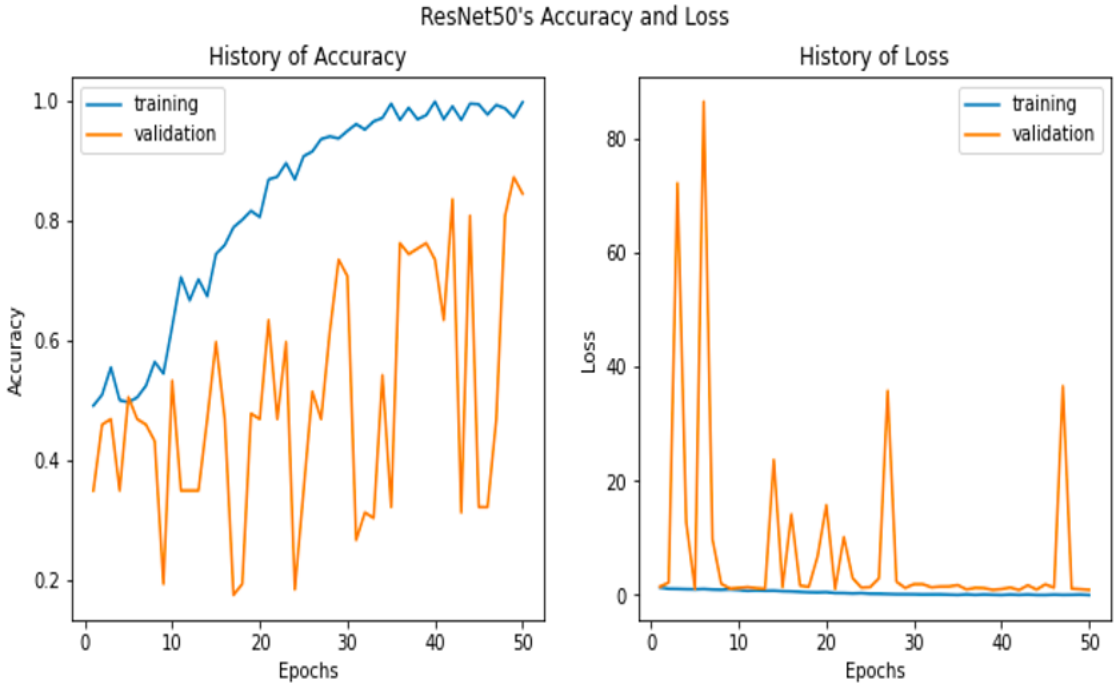

Figure 3: ResNet-50's Accuracy and Loss.

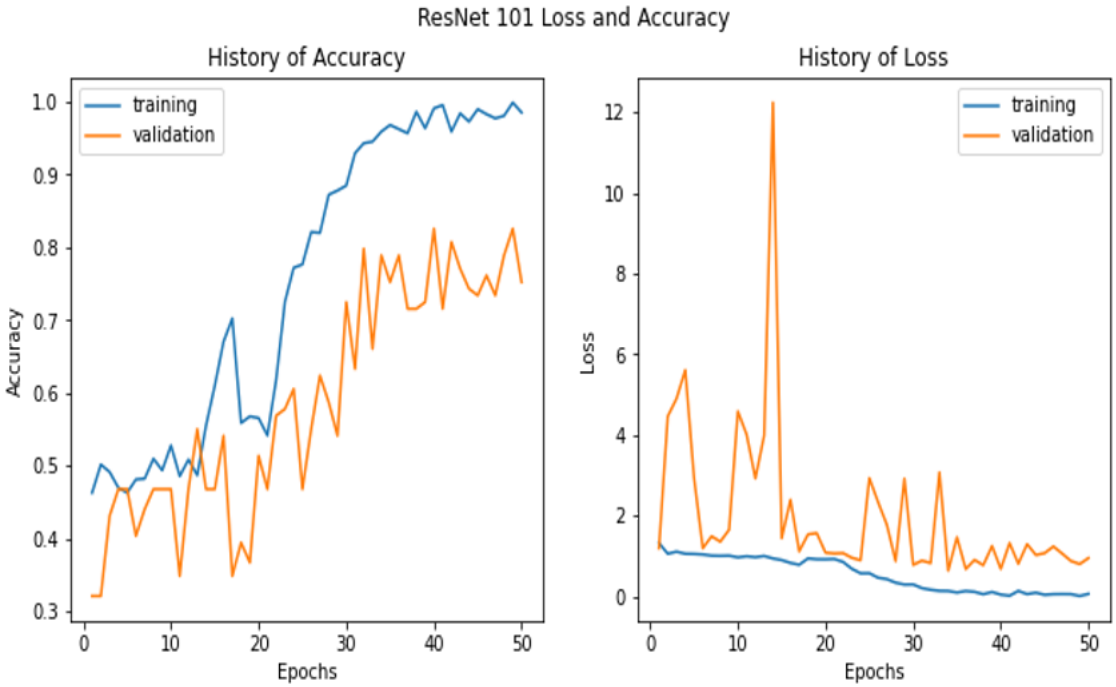

Figure 4: ResNet-101's Accuracy and Loss.

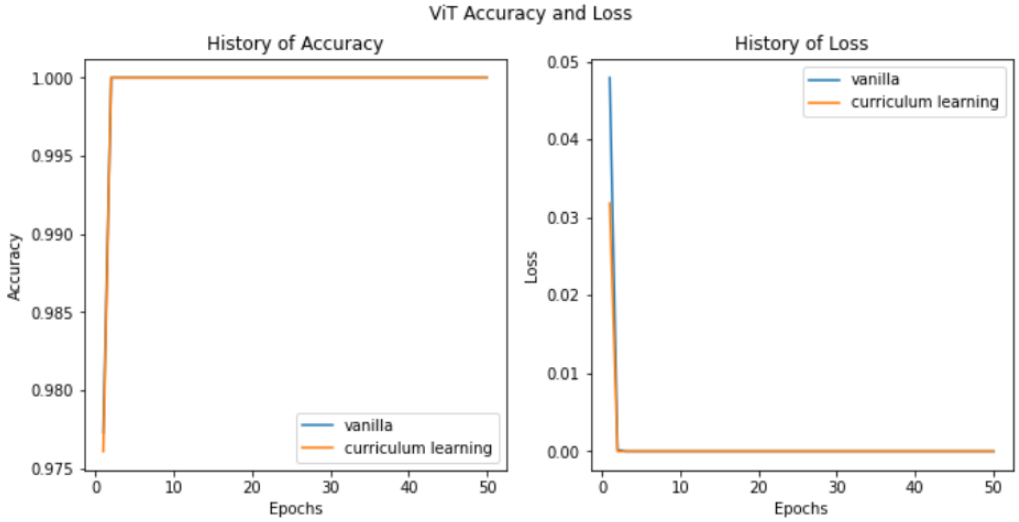

Figure 5: Vision Transformer's Accuracy and Loss.

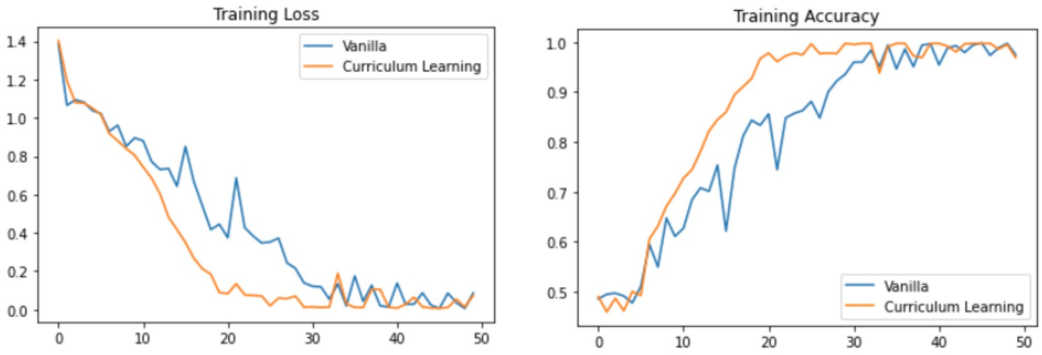

Figure 6: ResNet-50's Accuracy and Loss After Curriculum Learning.

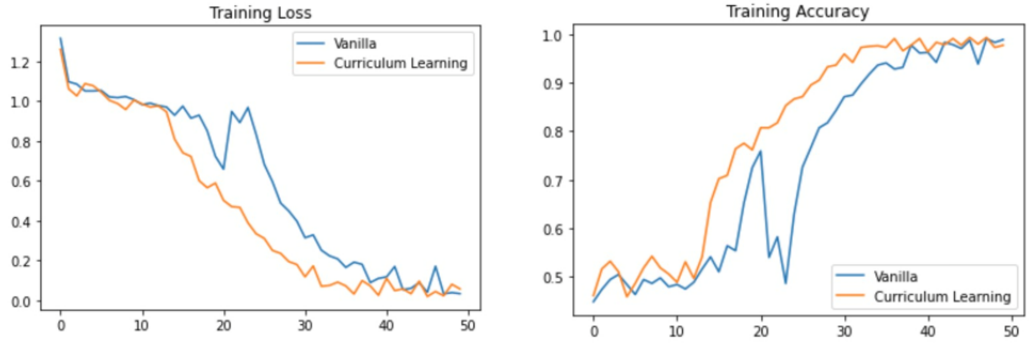

Figure 7: ResNet-101's Accuracy and Loss After Curriculum Learning.

6 Conclusion

In this study, we use an image classification dataset from the Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) to detect the tumor type using four distinct convoluted neural networks (CNNs). The approaches we utilized were VGG-16, Inception-V3, ResNet-50 and 101, and Vision Transformer (ViT). The model ViT achieves the highest accuracy of 100%. To boost model performance, we pre-organized our dataset using curriculum learning with Sparse Categorical Cross Entropy as the scoring function and a fixed batch size as the pacing function. The outcome demonstrates that curriculum learning significantly raises the testing accuracy of the models.

References

[1] "Key Statistics For Lung Cancer." *American Cancer Society*, 24 February 2022 https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html.

[2] "Lung Cancer Key Findings." *American Lung Association*, 17 November 2022, https://www.lung.org/research/state-of-lung-cancer/key-findings.

[3] Geng, Lei, et al. "Lung Segmentation Method with Dilated Convolution Based on VGG-16 Network." *Computer Assisted Surgery*, vol. 24, no. sup2, 2019, pp. 27–33. Crossref, https://doi.org/10.1080/24699322.2019.1649071.

[4] Guy Hacohen, Daphna Weinshall. (2019). On The Power of Curriculum Learning in Training Deep Networks. *International Conference on Machine Learning*, 2535–2544.

[5] Karen Simonyan, Andrew Zisserman. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Vision and Pattern Recognition*.

[6] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." *ArXiv.org*, 10 Dec. 2015, arxiv.org/abs/1512.03385.

[7] Dosovitskiy, Alexey, Lucas Beyer, et al.. "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale." *In International Conference on Learning Representations*, vol. 1, 2021.